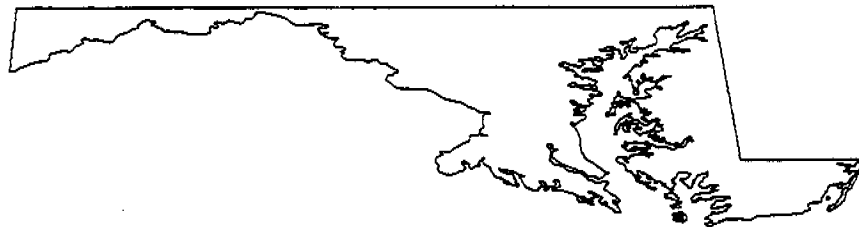


Designing End-of-Course High School Assessments

A Final Report to the Maryland State
Department of Education

Volume I



Wayne Camara
Howard Everson
Robert Majoros
The College Board

Kathleen O'Neill
John Fremer
James Braswell
James Carlson

Walter Jimenez
Ernest Kimmel
Patricia Klag
Michael Lapp
Timothy Ligget
Karen Nulton
Marlene Supernavage
Janet Waanders
Ann Marie Zolandz
Educational Testing Service

August 1997

In September 1996, the College Board was awarded the contract for the Design of the Maryland High School Assessment. Educational Testing Service (ETS) is serving as the subcontractor for this project. The work reported in this publication was supported by a contract from the Maryland State Department of Education.

The College Board is a national nonprofit association that champions educational excellence for all students through the ongoing collaboration of more than 3,000 member schools, colleges, universities, educational systems, and associations. The College Board promotes--by means of responsive forums, research, programs, and policy development--universal access to high standards of learning, equity of opportunity, and sufficient financial support so that every student is prepared for success in college and work.

Educational Testing Service (ETS) is a private, nonprofit corporation devoted to measurement and research, primarily in the field of education. ETS offers a wide range of products and services to improve education through the design, development, administration, and interpretation of high quality measurement tools. An extensive program of research on measurement theory, teaching and learning, and educational policy supports the programs and services provided to the educational community.

The College Entrance Examination Board is located at 45 Columbus Avenue, New York, NY 10023. Educational Testing Service is located on Rosedale Road, Princeton, NJ 08541

CONTENTS

	Page
Acknowledgments	3
I. Overview	5
II. Introduction	9
III. Test and Item Specifications	45
IV. Technical Specifications.....	61
V. Assessment Program Specifications.....	63
Appendix A – Additional test specification recommendations.....	A-1
Appendix B – Technical specifications	B-1

Acknowledgments

This is the second and final report by the College Board and Educational Testing Service to the Maryland State Department of Education concerning the proposed design of the High School Assessment. In January 1997, we provided the first report (referred throughout as “interim report”), which documented recommendations and alternative options regarding major design options for the tests and the public engagement process. We provided several presentations to the Maryland State Board of Education (MSBE) concerning these findings from Phase 1 of this study.

As with the earlier report, we would like to thank staff in the Division of Planning, Results, and Information Management (PRIM), especially Mark Moody and Steve Ferrara who worked with us extensively throughout all phases of this project. Dan Gadra was also instrumental in representing the concerns of various constituency groups as we developed specifications for the HSA. George Newberry served as the primary contact with MSDE during this phase of the project and was largely responsible for coordinating all aspects of the work completed since February on this design project. We are deeply grateful for his assistance, guidance, and direction. Gwenyth Swanson provided continued assistance in addressing the scoring and design of test items. Staff in Curriculum and Instruction, along with the co-chairs of the content teams for test specifications, were essential to the development of the test and program specifications. They provided important perspectives on educational, instructional, and administrative issues which will directly relate to the success of the HSA. Janet Bagsby, Gertrude Collier, Mary Jo Comer, Elaine Crawford, Gary Dunkleberger, Marci Emberger, Clarissa Evans, Barbara Graves, Cindy Hannon, Gary Hedges, Leslie Hobbs, Diane Householder, Diane Johnson, Linda Musial, Jessie Pollack, Robert Rice and others worked directly with content teams in reviewing major issues. Finally, we wish to thank all members of the various test specification committees, design team, and program specification committees who attending numerous meetings to derive recommendations on the design of the HSA.

I. OVERVIEW

This final report addresses issues identified in Phase 2 of the study and differs both in content and structure from the previous report. First, Phase 2 activities were primarily concerned with defining the actual test specifications for the 12 tests that were approved by MSBE in 1997. The test specifications can be considered preliminary “blue prints” for future forms of each test that will be developed by the test developer. We use the term “preliminary” because many specifications will need to be revised after pilot data are collected on the tests items that are actually developed and field tested.

The test specifications include such features as:

1. an overview of the design and features of all tests within a subject (e.g., math, science)
2. the types of items that will be included in each test form
3. subscores that *may* potentially be generated and used for indicating relative strengths and weaknesses
4. nontestable Core Learning Goals
5. links to Skills for Success Goals, where possible
6. recommendations from the Content Committees concerning specific aspects of the tests

The second chapter describes the processes used during Phase II of the project in developing the test and program specifications. It also reviews and summarizes major issues and findings on the overall design of the HSA. This chapter includes discussion of proficiency levels, design options, modules, local options for assessments, the use of HSA for higher education, accommodations, and other relevant issues. However, each of these issues is discussed in much more detail in the interim report, since these issues were addressed in the initial Phase of the HSA Design Project.

Chapter 3 of the report provides an overview of the test specifications that will guide test development. Additional, more detailed test specifications have been provided to MSDE that include the actual number of items, the specific links to Core Learning Goals, expectations, and in some instances, the indicators, and other details to guide test development. These detailed specifications have been classified as proprietary and are not for public release at this time. That is essential to ensure that students, parents, and teachers do not know which explicit indicator will be used for a 20-minute essay or which indicator will be measured by multiple items versus just a single item. Such information would be inappropriate to release now, or after tests are developed, because it would provide an unfair advantage to some students and encourage teaching to the test (over emphasis on some goals and ignoring other equally important goals in instruction and preparation). In addition, a few sample items are also contained in Appendix D. More information on the design of the tests, including sample items, will be forthcoming as test development proceeds.

Mathematics, science, and social studies have proposed use of the combination design for all tests, which will include approximately equal testing time for constructed response (CR) items (included limited and extended) and selected response (SR) items (multiple choice or machine scannable grid-in items). However, because CR items require substantially more time to complete than SR items, there will be proportionately more of the latter item types on all tests. The English test specification committee has proposed use of the Preparation Plus design, which will include a 60-minute preparation period, in addition to the 3-hours test length common across all tests. There are still some conflicts or remaining issues that must be resolved.

The proposed technical specifications for the HSA are briefly described in chapter 4, with additional detail in Appendix B. The validity, reliability, scaling, equating, scoring, and reporting of HSA results are addressed in this section. The reliability section discusses one conflict concerning the desire to have a high proportion of CR items to support instruction and educational reform and the minimum desired level of reliability recommended by MSDE (.90). A second important topic is the conflict that arises when the same test is used for both accountability functions (especially when multiple decision points are desired, such as with the HSA) and for information on relative strengths and weaknesses to support remediation. The same test is rarely effective for two such disparate uses, and the HSA is likely to be less effective for some of the proposed uses (e.g., accountability for students, accountability for districts and schools, diagnosis and remediation, instructional reform, placement or admission in higher education, certification of workplace skills or competencies). Other issues, such as the distribution of item difficulty, weighting various sections of the tests, test speededness, and the final proportion of CR and SR items permissible can only be addressed in broad discussions. Final test specifications cannot be developed on these issues until actual data are available from field testing.

Chapter 5 provides an overview of general test administrative issues concerning the operational implementation of the HSA. A number of strategic conflicts or disagreements among the various MSDE committees established during Phase 2 are addressed, and recommendations from CB/ETS concerning these issues (when they depart from committee recommendations) are presented. Generally, members of these committees advocated local choice on many administrative issues. However, as local choice and flexibility increase, less standardization can be maintained. Standardization is important if tests are designed for individual accountability because it can ensure all individuals are treated basically the same. Standardization is also a form of equity or fairness in terms of administrative time, conditions, comparability of scores—issues that are important if standard and fair consequences are to exist across students and schools. If there are standard consequences but different conditions (for administration, scoring, etc.), bias and inequity are introduced into a test score. These issues are addressed in terms of test administration schedules, breaks during the test, scoring, and test security. This section raises a number of significant issues and some conflicts that should be resolved soon. While there may be a desire to keep many “options” open, additional

decisions on the program specifications should be made soon to permit test development to proceed on schedule in an efficiency and cost effectiveness manner.

II. INTRODUCTION

1. Recommendations by MSBE

In February, the Maryland State Board of Education reviewed seven questions presented by MSDE, as well as the accompanying recommendations from MSDE. The State Board made the following recommendations concerning the HSA:

Question 1. *Will the high school assessment program design allow for individual student accountability?*

Recommendation: That the high school assessment program will allow for individual student accountability.

BOE Direction: To design the assessments to allow for individual student accountability in some form. If the Board should decide not to use the assessments for individual student accountability, the assessments could be used for individual school and school system accountability and for diagnostic purposes. It is possible to move from individual accountability to school and system accountability but we *cannot* move from school or school system level accountability to individual accountability.

Question 2. *Will the high school assessment program design allow for the program to become a requirement for graduation with a Maryland high school diploma?*

Recommendation: That the high school assessment program can be used as a requirement for graduation with a high school diploma.

BOE Direction: To design the test item specifications and thereby the high school improvement program to permit the Board to use the assessments for high stakes decisions which could include receipt of a Maryland high school diploma. The Board will consider the question of the relationship of the assessment program and the diploma later in 1997. Additional public engagement and public input is being solicited before the decision is made. (The State Board has modified its meeting agenda format to provide a 30 minute period of public comment relative the high school assessment program starting with the March 25-26, 1997 meeting.)

program. Content specialists and local school system staff are participating in the discussions.

Question 6. *Will the high school assessment program be designed to provide three annual test administrations with one makeup for each test?*

Recommendation: That it is acceptable to begin the program with two annual test administrations and one makeup for each administration.

BOE Direction: The design teams should develop sufficient specifications to allow for two annual test administrations and one makeup for each administration. The Board may authorize additional test administrations based on needs identified during the no-fault period. The Board asked staff to examine how the teaching of half-credit courses and the awarding of half-credits might be affected by program and psychometric considerations now guiding the development of the High School Assessment program. Staff is to define the extent of the problem and make recommendations for a solution.

Question 7. *Which design option(s) is (are) preferable?*

Recommendation: The Preparation Plus and the Combination designs are preferable.

BOE Direction: The design teams are to develop the Preparation Plus model for English, social studies, and science. The Combination model should be developed for mathematics. These are the design models recommended by the content teams. The Board recognized that staff has serious concerns about logistic and administrative problems with the Preparation Plus model. However, staff was directed to carefully examine possible solutions to these problems. It is possible to move from Preparation Plus to Combination once the design is complete. It is not possible to move from Combination to Preparation Plus. The Board may be asked to reconsider Preparation Plus.

2. An Overview Of Work Completed During Phase I Of The HSA Design

This section briefly summarizes the discussion and recommendations contained in the interim report to the Maryland State Board of Education produced by the College Board and Educational Testing Service (ETS) in January 1997 (Camara, Kimmel, et.al., 1997). For more

The preponderance of discussion and comments during the public engagement meetings focused on the first of these. Proponents of using the HSA for individual student accountability argued that:

- Students need to become responsible partners in their own education.
- High stakes for individual students will motivate parents and teachers in all grades to give greater emphasis to the development of the knowledge and skills outlined in the Core Learning Goals.
- Holding each student accountable is the best way of enhancing the value of the high school diploma.
- It would allow for verifying that the state has met its constitutional responsibility for ensuring the quality of the education received by all students.
- Expectations need to be raised for all students; this is best done by holding each student accountable.

The opponents of using the HSA for individual student accountability have argued that:

- The quality of education can be better improved through a focus on school accountability/ program improvement, as has been demonstrated by the use of MSPAP.
- Individual student accountability creates more practical and policy problems than it does incentives for improved learning. Problems include how to handle transferees, absentees, ESL and special needs students, and the logistics of record-keeping.
- Holding all students to the same high standards will result in an unacceptable failure rate across the state because of student variation and individual differences that are found in all indicators of student abilities and performance.
- There will be a disproportionate rate of failure among minority and/or poor students, creating major legal and political challenges to the HSA system, and/or resulting in the lowering of standards over time.
- The individual school/district is in a better position to judge whether a student meets graduation requirements, including, but not limited to, the Core Learning Goals.
- Preparing for the HSA tests will become the de facto curriculum, narrowing the scope of what is dealt with in each of the 12 relevant courses and the curricular and assessment options for local districts.
- The practice would not benefit high school students who have not been held accountable for the first eight or nine years of their schooling. It would also be unfair unless a level playing field exists across districts in terms of the quality of education provided to students.

Does MSBE want the HSA to continue to be used as an absolute requirement for high school graduation of all students in spite of the challenges to implementation and the opposition from key constituency groups?

Each of the three compensatory models creates a substantial counseling and course scheduling problem for the student and the school: if a student receives a below-satisfactory score on a test, should he or she receive remediation and re-take the test or assume that he or she will receive a higher, compensating, score on another test? Nonetheless, during public engagement activities, many educators voiced support for some form of a compensatory model. As noted above, it is more likely that a student will receive a diploma with the compensatory models than it is with the multiple hurdle option described as the first possibility.

A related issue in setting high standards for graduation using any of these models is “How High is High Enough?” The model selected by MSBE for the decision rule that will determine whether an individual student receives his or her diploma has substantial implications on the overall passing rate and difficulty of the HSA.

Which of these models best reflects MSBE’s intention for the decision rule to be used for the HSA graduation requirement?

2.4 *Implications of Ten Separate Tests*

The number of tests students must successfully complete is a significant issue in setting decision rules and determining the overall passing rate for students across the state and within each district. As noted above, requiring all students to pass 10 tests creates a substantial burden, but even compensatory models based on this number of tests (where students can combine scores across exams) will result in an assessment system that will be:

- overly complex and costly
- difficult to track and monitor student performance
- burdensome on local districts who must manage the administration and score reporting

Most high-stakes assessments require individuals to complete all assessments at one time. When scores are reported, the individual is aware of the outcome (pass or fail, or a specific score) and can make a decision about his or her future (e.g., the score is good enough and I will not retake the test; I will or must retake the test). As the number of separate tests required of students increases, the probability of failing one or more tests statistically increases as does the complexity and cost for managing the separate test administrations and the test score data and reports that must be issued following each administration. The detailed, elaborate, and costly procedures required to develop and maintain such an assessment system must be weighted against the advantages and disadvantages of requiring 10 separate tests.

Alternative designs that MSBE should consider (or reconsider) include:

- Developing end-of-program assessments in four areas.

throughout the year. This option proposes that such information be collected in a portfolio that would be scored by the classroom teacher. The information derived from the portfolio would be combined with a student's score on a timed (2-3 hour) assessment given under standardized conditions and scored centrally. This option provides for a wide sample of a student's work/performance to be used in evaluating achievement of the CLGs; it also allows for the classroom teacher's judgment to affect final pass/fail decisions.

2. Preparation Plus is based on the idea that students should share a common learning experience prior to assessment and that that learning experience can provide the substantive basis for a portion of the assessment. The common assignment or learning experience would model the kind of instructional strategies that are advocated for the particular subject. This common learning experience makes it possible for the timed (2-3 hour), standardized assessment to draw on more complex stimuli, e.g., a "messy" real-world mathematics application, than would otherwise be possible within the time limits of the assessment administration. The standardized assessment would be scored centrally.
3. Combination reflects a mature measurement technology that combines a number of constructed-response questions with a significant number of selected-response questions to create a reliable assessment instrument. Within the 2-3 hour time limit for administration of the assessment, it would include an extended constructed-response question that draws on a number of brief documents, laboratory results, or other authentic stimuli. The assessment would be scored centrally.
4. Limited Combination would be an entirely machine-scorable assessment administered in 2-3 hours in a standardized situation. Challenging selected-response questions would be the predominant type of question; although in the sciences and in mathematics, a machine-scorable constructed-response format would also be used. The answer sheets would be scored by MSDE.

PORTFOLIO PLUS	Combines information from student's portfolio, scored over time by the teacher, with information from an "on-demand" assessment including: <ul style="list-style-type: none"> • Two Extended Constructed Responses • Several Brief Constructed Responses • Selected-Response section
PREPARATION PLUS	All students are assigned a common learning task prior to the "on demand" assessment. This preparatory work is not scored but provides a common learning experience as the basis for a more specific task on the assessment. The assessment would include: <ul style="list-style-type: none"> • One Extended Constructed Response based on preparatory task • One Extended Constructed Response on another topic • Several Brief Constructed Responses Selected-Response section

program. Inclusion of performance assessments (associated with the first three designs) further extends the turnaround time and complexity for scoring the HSA. It takes several weeks to score high-stakes tests. The design options that include one or more extended constructed-response items, several shorter constructed-response items, and a selected-response section, will require approximately 7-9 weeks for score turnaround, and probably more time if Maryland teachers are to serve as scorers. The fourth option, which is entirely machine scorable, would require approximately 4-5 weeks or longer. It should be noted that meeting these estimates for score turnaround will require substantially increased financial and human resources from the state and the schools, as well as the successful resolution of a number of very critical issues.

Does MSBE want the HSA design to emphasize the importance of turnaround time for score reports at the expense of other design features?

A related issue is the timing of reporting scores to schools, students, and parents, e.g., must the scores be returned before the end of the semester or course? In planning the schedule for an assessment program, one can work backward by first specifying when the test scores must be reported and using that information to determine the latest date for administration. The high-stakes impact on individual students is cited as a major reason that scores must be returned by the end of the term. However, there are other reasons:

- To enable schools to plan remediation that may be required during the summer or in the following semester for students who fail, as well as to inform course placement decisions.
- To provide scores to seniors before graduation.
- To provide scores to students and schools while they have meaning. Quick turnaround is viewed as essential for ensuring that the assessments are viewed as important levers of reform in the schools.

Given the practical limits on turnaround time, there are two distinct alternatives¹:

- Administer the HSA part way through the course and report the results just prior to the end of the course (relaxing the association between assessments and conclusion of the course).
- Administer the HSA at the end of the course and report the results several weeks after the end of the semester.

These two alternatives are illustrated on the following page.

¹ These are the only alternatives, if MSBE retains the requirement to have ten separate end-of-course tests. Earlier discussions about moving to an end-of-program assessment system or introducing four content tests (which might be administered midway through the next course level) would provide additional alternatives that could address some of the concerns about score reporting.

2.9 *Flexibility In Accommodating Variations Among Local Districts*

The tradition in Maryland, as in most of the United States, of substantial local control of K-12 education makes the implementation of any centrally directed process difficult. HSA is no exception. There are persistent demands for choice and flexibility within the structure of the proposed HSA program.

The demands for flexibility relate to four central issues:

- Accommodating curricular patterns other than the twelve described by the Core Learning Goals reports.
- Accommodating half-credit courses.
- Accommodating accelerated courses.
- Accommodating districts and students by providing greater flexibility and choice through the use of modules.

2.10 *Alternative Curricular Patterns*

A basic principle underlying the conceptualization of any assessment program is that the assessments should be aligned with the curriculum in order to demonstrate curricular validity and to reinforce and support the standards implicit in the Core Learning Goals. However, many districts prefer assessments which conform to their specific course sequence, compositions, and instructional patterns - including integrated or half credit courses.

If MSBE wishes to accommodate districts employing or planning to introduce integrated courses or half-credit courses additional assessments are required. In math and science alone, there would be substantial additional costs and time required to develop and maintain these additional assessments.

The request for half credit assessments also raises severe logistical issues and psychometric concerns. Modules (or half tests) would not provide reliable scores for individual students. The combination of two modules, taken several months apart, into a total assessment score could not be interpreted in the same way as the total assessment taken at one time. The logistics of getting the appropriate half-tests to schools and students would add an expensive complication, while the need to create a system to match partial information from two separate half tests adds further expense. In addition, it is almost certain that schools that offer half-credit courses will not agree on how the content, skills, and processes from the entire course are distributed among their units. We believe all of these issues make administering and scoring separate modules for half-credit courses unfeasible.

Does MSBE want the HSA design to include specifications for integrated courses in mathematics and/or science (Integrated Mathematics I, Integrated Mathematics II, Integrated Chemistry/Physics, Environmental Science)?

There are additional difficulties raised with modules and other aspects of choice in a large-scale high-stakes assessment program such as the HSA. Each variation within a program, such as modules:

- introduces non-standardized conditions which make it difficult to make valid and fair comparisons of student and school performance.
- creates substantial logistical burdens for schools which must determine which variations of assessments are to be completed by which students and for the state (or its contractor) who must produce multiple iterations of each assessment and coordinate printing, shipping, and scoring across districts (the probability of errors which impact the quality of the results increases dramatically).
- significantly increases costs for development, implementation, psychometric studies, scaling, and scoring.
- complicates score reporting at the district and state level -- producing district or school reports that cannot be compared to each other.
- hinders students and teachers from gaining familiarity and comfort with the assessments and increases the difficulty of developing a common understanding of HSA by parents and the public.

Refer to extended discussion in the following section (see 4.2).

2.13 *Supporting Students Who Do Not Demonstrate Competence On The HSA*

A continuing issue is whether there should or can be alternative demonstrations of competence for students who do not pass one or more of the HSAs. Opinion during public engagement activities was sharply divided between those who believe that there should be local alternatives to passing the HSAs and those who argue that any local procedure is susceptible to manipulation and that the credibility of the HSA program will be undercut by permitting local alternatives to be used. The latter argue that such alternatives will be perceived as unfair.

Consultation with psychometric experts has indicated that there is no practical means of empirically demonstrating that alternative activities will provide “equivalent evidence of competence.” That is, you cannot align several different assessments to the Maryland state assessment and speak with any level of confidence about the equivalence of student performance across these different assessments (i.e., you can not statistically determine that students with a specific score on the local options would perform comparably to students with the same score on the state assessment in such an operational testing program).

If Maryland strongly desires to approve local alternatives for demonstrating competency on the Core Learning Goals, one possibility is to accept a lower standard for student comparability between these assessments. The state or local districts could develop one or more alternative options (or certify those options developed by LEAs) that can be used to demonstrate that standards have been met (by all students or just those who initially

Finally, very often, committees were unable to reach a consensus on an issue or did not have adequate time to address all issues presented. Individual members sometimes disagreed with the overall recommendation or expressed a different perspective on an issue. Some, but not all of this is captured in the TASC and PSC recommendations (Appendix E). The views and summary of recommendations which follows is the CB/ETS interpretation of a committee's recommendation based on materials (see Appendix E) available and our observations during meetings. Individual members of each committee may disagree with this summary as it concerns a specific issue.

MATRIX OF MAJOR ISSUES AND RECOMMENDATIONS

(Key to Committees: PSC = Program Specifications Committee, TASC = Test Administration Specifications Committee)

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION (when cell is blank CB/ETS support committee recommendation or have no comment)	ELABORATION IN SECTION __
Remaining Issues				
Local Scoring for CR items	MSDE conveyed HSA Task Force recommendation		Local scoring cannot replace MSDE-organized scoring (e.g., central, regional, dispersed)	II. 4.1; V. 10.4-10.5; Interim Report
Modules for Assessments			Modules are not appropriate or feasible for the purpose of the HSA	II. 4.2; Interim Report
Alternative Assessments			Alternative assessments can be used for students who fail the HSA if a judgmental process is used rather than requiring strict equivalence of competence	II.4.3; Interim Report
Standardized administrative and scoring procedures	TASC	Substantial flexibility should be provided to schools in developing administrative conditions, schedules for testing and security procedures	Standardized procedures are required for test administration, security, and scoring given the proposed uses of the HSA	II.4.4; Interim Report; App. E TASC Rec.; V 13.4
Multiple Uses of HSA			The same test cannot be equally effective for multiple purposes or uses and MSDE should prioritize proposed uses so tests can be designed (and evidence gathered) to support the most essential uses	II.4.4.1; Interim Report; App. B
Validity and Test Use			Validity is the overarching standard or criteria for evaluating the quality of the HSA. Given the intended purposes of the HSA, any feature which threatens the validity should be carefully considered by MSDE	II.4.4.3; Interim Report; App. B
Validity and Opportunity to Learn			MSDE should acquire evidence that curriculum and instructional practices across districts do properly include the CLGs through curriculum audits, teacher surveys and other means	App. B; Interim Report

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION __
Remediation	PSC	(1) Appropriate assistance must be available for all students. <i>If</i> a student does not wish the remediation and assistance available, parents should be required to sign a form (2) Appropriate remediation must be designed by local districts	MSDE must identify appropriate remedial strategies and work with local districts in supporting these activities prior to high stakes uses of the HSA	Interim Report
Test Security				
Guidelines for Security	TASC	MSDE should develop few new specifications; rather schools should have maximum flexibility in determine how to best implement security procedures	Detailed and standard security procedures should be developed centrally by MSDE	V. 1; App. E TASC Rec.
Student Identification	TASC	Student picture IDs should not be required of students completing the HSA	If picture IDs will not be required, alternative procedures should be developed by MSDE to ensure students can be correctly matched to tests forms and correct student identification can be verified.	V.1.1; App. E TASC Rec.
Student Identification	TASC	Require school staff to sign a from taking responsibility for verifying student names and identifications	Additional procedures are required and should be specified by MSDE	V.1.1; App. E TASC Rec.
Parental and Student Materials on Security	TASC	Several pamphlets should be developed by each district outlining all policies and inappropriate student behaviors for completion of tests. Students would then need to sign a form attesting that they received and understood policies and consequences	Standard materials should be developed by MSDE and not left to each district	V.1.2; App. E TASC Rec.
Descriptive Materials	TASC, PSC	Additional materials describing the purpose of the HSA should be developed and shared with parents, students (several years in advance of testing), teachers...		V.1.2; App. E TASC Rec. and PSC Rec.
Scrambling MC Items	TASC	Scrambling MC items across forms is desirable if it will increase security		V.1.3; App. E TASC Rec.

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION
Security Procedures	PSC	All materials must remain secure at all times	Detailed procedures similar to those developed for MSPAP must be developed to control for access and security of test materials prior to administration, during transition to scoring and score reporting, and after scoring is completed	V. 7.4; V. 1; App. E PSC Rec.
Disclosure of Forms	PSC	Test forms should be disclosed only after there are sufficiently large item pools which enable such release	Formal procedures and criteria must be developed by MSDE to guide this process. Once items are disclosed they should not be used on a test form	V. 7.5; App. E PSC Rec.; Interim Report
Review of Materials or Items or Student Responses	PSC	(1) Any review should be conducted under secure conditions, (2) If parents or citizens are to have access to test materials, there should be defined procedures in place, (3) Copying of test items should be prohibited unless the item is disclosed		V. 7.4; App E. PSC Rec.
Test Administration				
Manuals	TASC	Defer development of manual and use MSPAP as the primary guide	MSDE should review existing manuals used in other state testing programs and develop a manual for the HSA prior to the no-fault testing. Procedures used in other state testing programs must be reviewed for appropriateness with HSA uses	V. 2
Breaks During Test Admin.	TASC	Recommend 1-2 breaks with districts and schools making their own determination and establishing their own time limits for breaks	Support recommendation from PSC; number of breaks cannot be a local choice because of security and fairness issues	V.2.1; App E TASC Rec.
	PSC	Recommend 0-1 breaks with MSDE making determination		V.2.1; App E PSC Rec.
Test Length	MSBE	3 hours	Support recommendation from MSBE and PSC for standard administration times of no more than 3 hours. Tests can be designed to reduce speededness but variable time cannot be accommodated because of proposed uses.	V.1; V.2.2;

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION (when cell is blank CB/ETS support committee recommendation or have no comment)	ELABORATION IN SECTION
Test Length	TASC	Variable time - allow all students to complete the tests	Support recommendation from MSBE and PSC for standard administration times of no more than 3 hours. Tests can be designed to reduce speededness but variable time cannot be accommodated because of proposed uses	App E TASC Rec.
	PSC	All tests should have the same time limit		App E PSC Rec.
Test Grouping (Setting)	TASC	Permit each school maximum flexibility to administer tests in large settings, classes or other appropriate environments	While schools should determine where tests can best be administered, MSDE should consider some general guidelines	V.2.3
Eligibility for Proctors			Employ criteria used for MSPAP	V.2.4
Training for Proctors	TASC	LACs determine the appropriate procedures for training proctors at their schools and develop materials	Standard materials should be developed by MSPAP, not each LAC. Training should be conducted as part of the state-wide field tests	V.2.4
Number of Test Administrations	MSBE	Begin HSA with 2 annual administrations		
	TASC	Support MSBE, but ask for a rationale to provide districts with half-credit courses and quarter system		V.2.5; App E TASC Rec.
	PSC	Recommend a third summer administration		V.2.5; App E PSC Rec.
Completing 2 tests from Same Content Area in the Same Test Administration	TASC	Students should be permitted to complete two tests from same content area (e.g., mathematics) during same testing period (spring 1999)		V.3.1; App E TASC Rec.
Students Completing a Test when they have <i>not</i> Enrollment in the Corresponding MD Course	PSC	Permit districts to determine their own policy	MSDE should establish a standard policy	V.3.2; App E PSC Rec.
Out-of-State Students	TASC	Permit students to take HSA if merit or preferential admissions associated with higher education uses	Consider all such extended uses of HSA after initial implementation and operation of core program	V.3.3; App E TASC Rec.
Private School, Hospital-bound and Home-bound Students	TASC	Should not be part of the HSA unless they are to receive a MD diploma		V.3.3; App E TASC Rec.

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION
Out-of-State Transfer Students	TASC	Require students to complete any test that corresponds to a course they must complete in a MD school	Support TASC recommendation. MSDE should determine what % of a course (or instructional days) must be completed for mid-year transfers to be required to complete test	V.3.4; App E TASC Rec.
In-state Transfer Students	TASC	Students must complete any HSA test they have not already passed		V.3.4; App E TASC Rec.
Exemptions for AP exams	TASC	No exemptions	Reconsider decision based on probably negative consequences for high achieving students. Conduct a concordance or validity student to determine if AP performance exceeds MD standards in subjects where AP is offered	V.3.5; App E TASC Rec.
Exemptions for IB	TASC	No exemptions	Reconsider decision based on probably negative consequences for high achieving students.	V.3.5; App E TASC Rec.
Accommodations for Students with Disabilities				
Develop guidelines	TASC	Defer action	Do not defer beyond this winter since policies must be in-place and communicated well in advance of testing	V.3.6; App E TASC Rec.
Accommodations	TASC	Accommodations used for test must be the same as those used for instruction		V.3.6; App E TASC Rec.
Familiarity with Accommodations	TASC	Students must receive the same accommodations in instruction prior to testing to ensure they are familiar and comfortable with accommodations		V.3.6; App E TASC Rec.
Meeting of Special Ed. Directors	TASC	Convene a meeting to resolve remaining issues	MSDE, with input from special education directors, should resolve these issues	V.3.6; App E TASC Rec.
Braille forms	TASC	Braille forms should be developed where feasible	At least 80% of each test should be designed so that it can be transformed to Braille	V.3.6; App E TASC Rec.
Temporary Disabilities	PSC	Use same procedures as used with students having permanent disabilities		App E PSC Rec.
Record of Accommodations and Flagged scores	PSC	(1) Scores should not be flagged (2) A record of accommodations should be maintained in student file		V.3.6; App E PSC Rec.

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION
Large print forms			Large print forms should be produced for all tests	V.3.6
Computer delivery of tests	PSC	Should be introduced as soon as feasible	Requires further examination for determining equivalence and costs and equipment needs may be barriers for several more years	V.3.6; App E PSC Rec.; Interim Report
LOEP Students				
Scoring CR Items	PSC	Special scores and/or special training may be required for LOEP students	Requires further examination	V.3.7; App E PSC Rec.
Meeting of LOEP Directors	TASC	Convene a meeting to resolve remaining issues	MSDE, with input from directors, should resolve these issues	V.3.7; App E TASC Rec.
Scheduling				
Date of Testing	PSC, TASC	All students across the state will take the same HSA tests on the same date(s)		V.4.1; App E TASC and PSC Recs; Interim Report
Time of Testing	TASC	Provide schools with some flexibility when test begins and ends	Support recommendation as long as the variance of starting times is less than 2.5 hours	V.4.1; App E TASC Rec.
Extending Administration of 1 Test Over 2 Days	English and Math Content Teams	Recommend 2 days for each of their exams	Disagree with English and Math Content Teams and support the TASC recommendation	V.4.2; App E TASC Rec.; V.1; V. 5.2; App B
	TASC	All tests should be administered entirely on one day		
Instruction on Testing Dates	TASC	Recommend each district determine if instruction should be conducted on testing days or if classes should be canceled		V.4.3; Interim Report; App E TASC Rec.
Preparation Plus				
New model of Prep Plus	English Content Team	Provide all students with 60-minutes of instruction on the day of testing. Preparation activity is highly scripted.	Support the new model which does significantly reduce technical concerns. Support TASC recommendation for feasibility study	V.5

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION
New model of Prep Plus	TASC	(1) New model should be approved by MSBE. (2) A state-wide feasibility study should be conducted with actual master school schedules prior to implementation. (3) There is no consensus on whether this model can be successfully implemented in the schools.	Support the new model which does significantly reduce technical concerns. Support TASC recommendation for feasibility study	V.5; App E TASC Rec.
Implementation of Prep Plus	English Content Team	Appears to disagree somewhat with CB/ETS recommendation	Substantial operational and administrative problems have not yet been resolved and further examination is required. This model will place more burden on schools, increase overall testing time, and increase the cost of the HSA	V.5; Interim Report
Teachers for Preparation	TASC	MSDE should investigate whether adequate numbers of certified teachers will be available to conduct preparation activities	This may or may not be a problem depending on the scheduling selected for the HSA and the level of certification required for proctors (must they be certified in English or can other teachers administer these tests?).	V. 5; Interim Report; App. E TSC Rec.
	English Content Team	Believes their recommendations, if implemented by MSDE, will eliminate these concerns		V. 5
Equipment				
Calculator Use on Math Tests	Math Content Team	Require a graphing calculator for all math assessments	Staff must have calculators and have adequate training and continuing support for instruction well before calculators are required for proficiency on tests. Given the current time table MSDE should either delay test schedule or phase in use	V. 6; App. G
Staff Development in use of Graphing Calculators	TASC	Teachers will need staff development in the use of graphing calculators	A survey of teachers should be conducted this fall to determine teachers' current and future use of calculators and competency in the use of calculators	App. G

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION __
Rationale for a Graphing Calculator	TASC	In the CLGs	Implementation of this CLG may be several years away, and students must have an opportunity to learn prior to requiring calculator use for assessments. MSDE must determine if construct irrelevant variance will be introduced with this requirement before designing tests to require graphing calculators	App. G, Interim Report
Pilot Test	TASC	Before implementation, a pilot test should be conducted to ensure type, brand, and functions on calculator, and previous familiarity and use of calculator effects on performance		App. G; App E TASC Rec.
Calculator Functions	TASC	MSDE should specify the exact functions calculators should and should not possess	MSDE should consider specifying the actual brand and models which meet their specifications and limit calculators to a few such models	App. G; App E TASC Rec.
Calculator purchases	TASC	State and local districts must purchase adequate numbers of calculators required for testing, not just class instruction. In addition, MSDE should form a committee to develop a request for funding of this equipment by the General Assembly		App. G; App E TASC Rec.
Dictionaries	Subject Content Teams	Dictionaries should be available to students throughout the testing where permitted. Dictionaries should not be accessible for science tests.		III 2.7 and 2.8.2
Technology	PSC	Recommends word processors, Internet, and computers be employed for HSAs as soon as feasible	Requires further examination. Rationale for requiring technology in assessments differs from rationale for technology in instruction	V. 8; App. E PSC Rec.
Manipulatives in Math Test I	Math Content Team	Items should generally not require manipulatives		App. C
Test Development				
Response Choices on MC Items	PSC	All MC items should have 4 answer choices with only 1 correct response		V.7.1; III 2.3

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION ____
Response Choices on MC Items	PSC and all Subject Content Teams	Answer choices and distracters should not include Roman numerals, "all of the above" and "none of the above" choices		III 2.3
Test Content Involving Topics Portraying Minority Interests	PSC	Recommends that about 33% of all items portray minority topics	Suggest using a 15% minimum for minority topics, while striving for a higher portion of items if feasible and cost effective	III 2.8.1; V. 7.2
MC Item Blocks			Arrange MC items in blocks so students will not be required to switch item formats and to reduce costs	III 2.4
Visuals			Limit the number of visuals printed in half-tone or color to reduce costs and logistical problems	III 2.5
Inappropriate content	Subject Content Teams	Sports scenarios should not be used		III 2.6
Sensitivity Review	PSC	All individuals should be knowledgeable in relevant CLGs		App E, PSC Rec.
Subscores	Subject Content Teams	Subscores should be produced where feasible and as recommended by each content team		III 3.3, 4.3, 5.3, and 6.3
Social Studies and MC Items	Social Studies Content Team	Requests exemption from a technical specification - Reduce the # of MC items required on all tests from 90 to 60 items	Believe this will not be sufficient to ensure adequate reliability. MSDE should conduct a pilot study or modeling before any test assembly work is conducted with fewer than 90 MC items	III. 6
Preparation Materials	Social Studies and Science Content Teams	MSDE should develop a packet of test preparatory materials for teachers to use including suggestions on teaching key skills, model activities for CLGs, sample items and scoring rubrics	Recommend this be applied for all content areas	III. 6.1
Staff Development				
Staff Development	TASC, Content Teams	Supported CB/ETS recommendation	Extensive staff development is required prior to implementation of HSA for: (1) operationalization of CLGs, (2) test familiarization, instruction, and remediation, and (3) test administration	V. 9, Interim Report

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION
Scoring and Score Reports				
Appeals	PSC	MSDE should establish a defined appeal process for students and parents wishing to question a test score		V. 10.1
Proficiency Levels	PSC	Should be uses in scoring and score reports	Fewer decision points and proficiency levels are preferred given proposed uses of HSA and other constraints	V. 10.2; App. B; Interim Report
Scheduling of Testing and Score Reporting			MSDE must come to closure on the approximate calendar for testing and whether scores will be provided prior to, or after the end of the semester	Interim Report; V 10.3
Number of Readers for CR Items			A smaller number of readers is preferable to a larger number of readers, given other constraints for score reporting. All CR items should be scored by more than one reader and interreliability estimates should be computed	V 10.4; Interim Report, App. B 5.1
Not Reached Items	Subject Content Teams	Do not count as incorrect responses (Math team did not discuss this issue)		App. B 7.2.1
Scoring Range for CR Items			A limited scoring range (e.g., 1-5) should be used on all extended CR items	App. B 7.3
Field Testing & Quality Control				
Item Development			At least twice as many items as needed for operational forms should be developed	App. B 5
Field Testing MC Items			MC items should be field tested on operational forms	App. B 5 and 5.1; Interim Report
Field Testing CR Items			CR items should be field tested on samples recruited out-of-state	App. B 5 & 5.2; Interim Report
Field Testing for Psychometric Specifications			Initial field testing should be conducted prior to test development of operational forms to ensure the proposed specifications (e.g., items, reliability) are adequate and do not need modification	App. B 5; Interim Report

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION __
Quality Control Procedures			MSDE should establish a comprehensive set of quality control procedures which include audits of test development, pretesting, administration, scoring and score reporting	App. B 5; Interim Report
Database management	TASC	A minimum of one dedicated staff person will be required at each school	MSDE will require substantial additional staff support and appropriate information management technology to monitor student completion, retesting, and remediation for HSA	Interim Report; App. E TASC Rec.
Analyses for Detection of Cheating			MSDE should determine if routine analyses will be incorporated for detection of cheating	App. F; App B 9.4
Reuse of MC Items			Do not use the same MC item on more than 4 operational forms, spaced out in administration time	App. B 12
Reuse of CR Items			Use items once on an operational form only and then again, if needed, on one make-up form of the test, spaced out in administrative time if feasible	App. B 12
Psychometrics				
Content Validity			Each test form must be demonstrably related to the skills and competencies in the CLGs	App. B 2.1
Minimum Reliability			Overall reliability for each HSA test should be .90 or higher given the proposed uses of the HSA	App. B 3.1
Calibration Samples			A minimum of 1000 students should be used in the calibration of MC items and 1500 students for CR items	App. B 5.5
DIF			MSDE should incorporate DIF analyses in test development and eliminate items with extreme DIF	App. B 9.2

MAJOR ISSUE	MSDE COMMITTEE	MSDE COMMITTEE RECOMMENDATION	CB/ETS RECOMMENDATION <i>(when cell is blank CB/ETS support committee recommendation or have no comment)</i>	ELABORATION IN SECTION ____
Special Studies	TASC	A number of special studies should be designed during initial field testing to examine the most challenging and innovative aspects of the proposed design (e.g., calculator use, Preparation plus, MC/CR ratio of items, speededness) prior to development of operational forms	Strongly support recommendation	V. 12; App. E TASC Rec
Transition to HSA				
Number of Tests Introduced			If financial, administrative, or technical constraints remain it is better to introduce fewer tests than all 12 tests initially - phase in tests	V. 13.1
Reconsider justification for chemistry and physics tests	Science Content Team and TASC	Disagree with this recommendation	Reconsider justification for chemistry and physics tests given projected volumes, costs, and educational value	V. 13.2; Interim Report
Phase in High Stakes			If forced with a choice, phase in the high stakes but begin with the high standards. Changing standards will create uncertainty about the level of competence required of students	V. 13.3; Interim Report

4. Unresolved Issues

Some of these issues remain unresolved. If the Maryland High School Assessment Program (HSA) is to be implemented so that all students are held accountable to a common standard and face common and fair consequences in relation to performance on the set of assessments, there must be a strong degree of commonality or standardization in the content of the assessments, the conditions under which the assessments are administered, the process by which the constructed-response questions are scored and combined with the results of the selected response sections of each assessment, and the way in which scores and distributions of scores are reported and interpreted. This need for a standardized approach to assessment is incongruent with the desires of many in Maryland for the assessments to be more reflective of or adaptable to the conditions of particular school districts and to be usable for purposes other than certifying the achievement of competence on the Core Learning Goals. To the extent that local choice and flexibility increase, standardization can no longer be maintained -- thus raising questions of the fundamental comparability of the scores and the fairness of applying a common standard to all students. Conversely, if the Maryland State Board of Education were to decide that individual students will not be held accountable nor experience significant consequences for poor performance, greater flexibility and local choice could be included in the design of the program. These areas of disagreement were discussed at length in the January 1997 Interim Report to MSBE. The basic considerations are recapitulated here because of the persistence of several areas of tension.

4.1 *Local Scoring*

Some have argued that classroom teachers or the local school system should score the HSA constructed response questions in order to obtain the results in a more timely fashion than will be possible with a centralized scoring process and/or to ensure that more teachers have the experience of participating in the scoring process. Although some have argued that local scoring could replace a MSDE-organized scoring, most have argued that local scoring would precede a centralized scoring or centralized audit process.

The disadvantages of local scoring fall into three categories:

Economic/Operational: Superintendents and other administrators have expressed their concern about the expense of doing a local scoring and the impact it would have upon the time of the teaching staff. Teacher representatives have indicated that it would not be acceptable for scoring to be an additional, uncompensated task. The coordination and monitoring of dispersed classroom teachers scoring some number of assessments is daunting; the logistics of planning and supervising a system-wide scoring session which brings the teachers together is very challenging. Districts that have held scoring sessions have found them to be very difficult and expensive. In addition, scoring the selected response questions and matching the results with the constructed response results,

transforming the totals to a common reporting scale, producing reports, etc. all require skills and resources that most districts do not have.

Technical: Because important decisions will be made about individual students, it is critical that the assessment results be highly reliable and based on common performance standards. Getting a group of teachers (or others) to score constructed response answers consistently over students and over time requires a great deal of training, monitoring, and retraining. In order for the scores in different districts to be comparable, a cadre of teacher-leaders from each district would need to be thoroughly trained and prepared to train their colleagues in the district.

Public Perception: If local scores are used to meet graduation requirements and even if the technical and economic/operational challenges could be overcome, it seems likely that students, parents, and the general public would question the fairness with which the “common” standards had been implemented in their or a neighboring district. Legal or administrative challenges can be anticipated, requiring difficult and expensive demonstrations of the comparability of scores deriving from different districts.

Alternatively, if the MSDE scores the student responses centrally after they have been scored at the district level, there will be protests and challenges to the disparity between the scores given through the two different processes.

See pp. 46–48 of the January Interim Report for further discussion of these issues.

4.2 Modular Test Design

Since the earliest discussions of HSA, educators throughout Maryland have advocated that the assessments be constructed in modules. However, widely varying purposes have been cited as the reason to modularize the assessments. The purposes cited include:

- to have the assessments match integrated mathematics and science courses,
- to facilitate the testing of students enrolled in half-credit courses,
- to allow districts to pick and choose from a pool of modules those that most closely reflect local curricula,
- to allow districts to combine HSA modules with local assessments,
- to accommodate students with disabilities who need to complete the assessment in close proximity to instruction.

Although test modules can be used successfully for certain assessment purposes, they are not appropriate for the proposed use of the HSA to make high-stakes decisions about individual students.

There are three types of difficulties associated with the use of modules in HSA:

Comparability: Permitting districts to select among several modules would result in students in different districts completing different modules with no way to compare individual student-level performance with adequate precision to justify high-stakes (e.g., denial of state-endorsed diploma) uses with students. Variability among students' performance across districts may be as much attributable to the modules selected as it is to students' competency of the Core Learning Goals. The use of modules violates the assumption of standardized conditions that are required to make valid and fair comparisons of student and school performance. The complexities of equating and scaling the various combinations of modules make it virtually impossible for the scores to have the psychometric and technical qualities needed to make valid and comparable inferences about which students have met the required level of performance in terms of the Core Learning Goals.

Logistical/Costs: Modules would create substantial logistical burdens for schools to determine which variations of assessments are to be completed by which students. Similarly, the logistical burden on MSDE or its contractor to produce multiple variations of each assessment and to coordinate printing, shipping, and scoring, and score reporting dramatically increases the probability of errors that will impact the quality of the results. Such complexities significantly increase the costs for development, implementation, psychometric analyses, scaling, scoring, and standard setting.

Public Perception: The use of modules would hinder students, teachers, parents, and the general public in developing a common understanding of the Core Learning Goals and the related HSA assessments. Further, the use of modules would create an impression of unfairness. This impression will be exacerbated by the difficulty of empirically demonstrating the equivalence of different combinations of modules. The public will have difficulty understanding how the performance of students who have taken different tests, i.e., different combinations of modules, can be compared. As with any perception of unfairness, schools and the state will be vulnerable to administrative and legal challenges by those who believe they were wronged.

See pp. 27-28 and 124-127 of the January Interim Report for further discussion of these issues.

4.3 Alternative Assessments

There continues to be ambivalence within Maryland as to whether the results of alternative assessments or other forms of evidence should be accepted in lieu of HSA results, especially for students who do not succeed initially on the HSA. Some believe that local alternatives to passing the HSAs are needed while others argue that any local procedure is susceptible to manipulation and that the use of alternative assessments will undercut the credibility of the HSA program and the high school graduation requirement.

As with so many HSA design decisions, the technical feasibility of any alternative assessments depends on the use that will be made of the results.

If the results are to be used to make comparisons among students or to compare student performance across schools or districts, the same psychometric difficulties are present as in the use of different modules --except possibly in a more severe form. Many of the logistical, administrative, and scoring issues are similar also. In brief, there is no practical means of empirically demonstrating that alternative assessments will provide equivalent evidence of competence in reference to the Core Learning Goals. That is, you cannot align several different assessments to the HSA and speak with any confidence about the equivalence of student performance across these different assessments, i.e., one cannot statistically determine that students with a given performance on an alternative assessment have performed comparably to students with a particular score on the HSA. Yet such equivalence is required by professional standards and would appear to be essential both from a legal and an ethical perspective if the results are to be used for high-stakes decisions.

If the state and districts were willing to forego making any comparisons among students, schools, or districts, one can conceive of a judgmental process whereby experts would judge that various alternate assessments provided acceptable (but not equivalent) documentation of competence on the Core Learning Goals. Such a process would provide evidence that a student had demonstrated competence in Core Learning Goals, but it would be a very shaky basis for making a high-stakes decision such as withholding a diploma.

If the MSBE determines that it is necessary to be able to compare students or groups of students, the acceptance of alternative assessment options would add ambiguity to the system because the equivalence of the results cannot be demonstrated, but considerable expense will be generated in trying to do so. If comparisons and high-stakes individual decisions are not going to occur, alternative assessment options can be explored further. This path has many unknowns and is likely to prove both costly and lengthy.

See pp. 29-30 and 127-129 of the January Interim Report for further discussion of these issues.

4.4 Standardization versus Local Adaptations: Implications for Fairness

During the second phase of the project, CB/ETS has worked closely with Maryland educators who served on various committees and work groups advising MSDE on the test and program specifications. Many features valued and desired by some educators include:

- local adaptability and flexibility in test content and coverage of Core Learning Goals
- tests comprised of modules to permit variations according to local curriculum (when courses are offered, how goals are covered in various courses, when courses are offered, flexibility for semester or quarter courses)

- assessment that reflects student growth over time as opposed to on-demand testing
- assessments embedded in instruction (editing work, selecting among best pieces)
- flexibility (e.g., variations) in administration and scoring to accommodate local needs
- minimal state mandates in terms of content, administration, and scoring
- teacher scoring of student responses

However, while such features are preferences of some, it appears that few districts actually offer curricular that would benefit from a modular assessment approach, offer half-credit courses, or are reluctant to offer the test on the same dates and under the same administrative conditions according to MSDE staff serving on the design team.

Flexibility and adaptability can be accommodated somewhat more easily in tests designed for low levels of accountability (diagnostic uses) than they can when comparable performance across students and schools are required for high-stakes uses. The desire for greater flexibility and less standardization appears to conflict with the proposed use of the HSA as a graduation requirement. The preferences above cannot be easily accommodated in a high stakes individual assessment program. Similarly, the intended purposes of the HSA (e.g., certification for student graduation, district and school accountability) may require a level of standardization that is undesirable to some educators.

CB/ETS have assisted MSDE in developing test specifications with sufficient detail to ensure fairness of the resulting HSA tests. Willingham and Cole (1997) propose three criteria for evaluating fairness of tests:

- how useful the test is in serving its intended function
- how fair the test is for individuals and groups of examinees
- how well the test meets practical constraints

4.4.1 Fairness in Test Use

As noted throughout this report, a test cannot be equally useful for the variety of purposes proposed and implied for the HSA. That is, the HSA may not be equally useful in serving as an individual accountability measure for students and schools, as a lever of reform and educational and instructional change, as a standard for graduation and college admissions, and as a diagnostic instrument for students and educators. Some of these uses will be served well by the test, while other uses will suffer and be more problematic or questionable. There is no test that can be used for all these purposes (see appendix B, 3.3). When multiple uses are proposed for a single test or assessment, the probable result is that the test will be misused in some instances and negative consequences will result.

4.4.2 Fairness for Individuals and Groups

Fairness can be examined from the perspective of individual students and groups of students. As noted in Willingham and Cole (1997, p. 183), “the most useful test for all

high school physics students would include a careful sampling of all knowledge and skills typically included in the course, but that would not be the fairest test in schools that had no physics laboratory. On the other hand, reducing test content to what is available in every school threatens the usefulness and fairness of the test for students generally.” The same illustration applies to differences among schools in the structure of their curriculum (and coverage of Core Learning Goals, use of equipment such as graphing calculators, and preparatory activities). Issues concerning individual fairness arise when variations in the test environment, administration, scoring, and test content can have a substantial impact on differences in student performance. Subtle content differences between forms may have only a small impact within group analyses, but can bias comparisons across groups (Dwyer, 1979). The same principles apply to subtle differences in administrative conditions and scoring. While total standardization of all procedures is not possible with any test, lack of standardization will introduce unfairness in the process.

MSDE has developed a common set of Core Learning Goals and an HSA design that extends across all schools. Similarly, a standard set of consequences will result from the same test scores for all students and schools in Maryland. Differences in administrative conditions, test content, and scoring are inconsistent with decisions and processes already in place for one common assessment system, one set of content goals, and one set of consequences for test use.

4.4.3 Fairness and Practical Restraints

Proposed administrative specifications attempt to ensure that the HSA will be administratively feasible across all schools. Administrative conditions that may be maximally convenient to a high school in Prince George’s County may be unfeasible to a high school in Frederick County. Simply permitting schools or districts to choose administrative, scoring, or content specifications that are most efficient for them introduces a lack of standardization that threatens the validity of the tests. Common administrative specifications may not be maximally efficient for any school (or desirable from an instructional or local perspective), but they should not be totally unfeasible for any school. In addition, these proposed specifications will provide a common blueprint that provides an appropriate level of comparability.

Although usefulness, fairness, and practicality are essential features of a successful test, they are often viewed by some educators as social criteria employed by test users. Validity is always the overarching standard of test quality. It is highly desirable that information from the HSA be viewed as useful to policymakers, and the tests should be viewed as practical by administrators and educationally relevant by educators. But if the tests then lack validity, proposed uses cannot be supported.

4.4.4 Conflicting Desires for local flexibility and adaptations versus high stakes individual uses for the HSA

Many of the features desired by educators may not be in the proposed final design because the tests must be developed to permit individual high-stakes uses. Ultimately, any test design is a compromise between competing content goals, item types, educational preferences and psychometric requirements, and many other pressures. The HSA design is no exception—it too represents compromises. The resulting proposed specifications represent a responsible compromise to support the unique performance assessment movement ongoing in Maryland while maintaining adequate psychometric rigor and administrative feasibility required for the HSA intended uses. Some specifications will need to be reconsidered once test data are available. Other remaining conflicts must be decided, and a final finished direction for the administration and scoring of the HSA should be produced and shared with students, parents, educators, and Maryland citizens in the coming two years.

Finally, other issues of importance concerning the HSA have been discussed in the Interim Report to MSBE and are not repeated in this report. The feasibility of modules, local scoring options, the use of HSA for higher education, and the use of technology are examples of issues discussed at length in that earlier report.

III. TEST AND ITEM SPECIFICATIONS

The following chapter includes general discussion of the content specifications for the twelve HSA tests. The full set of content specifications is contained in Appendix C. Throughout this document, the terms “test specifications” and “item specifications” should be considered as proposed specifications.

1. Overview and Organization

1.1 *Structure of Test Specifications*

Test specifications serve as the blueprint for the test; they provide information that item-writers use in developing items and they provide instructions to test assemblers about how to select items for the test. These test specifications serve to define as clearly as possible the scope and emphasis of the test. In the case of HSA tests, they relate the test content to the Core Learning Goals (CLG) by identifying the relative importance of each section of the CLGs, they provide information about the nature and distribution of the item types to be included in the test (e.g., selected response or constructed response), and they indicate specific item-level constraints (e.g., sports contexts for test questions). By ensuring that the tests adequately cover the CLGs and are directly related to the content and skills that students are expected to know, the content validity of the HSA tests will be strengthened.

1.2 *Groups Involved in Setting Test Specifications*

In order to develop appropriate specifications for the tests and test administration, MSDE recruited members for three types of committees: four Test Specifications Committees (English, Mathematics, Science, and Social Studies), one Program Specifications Committee (PSC), and one Test Administration Specifications Committee (TASC). Members on all committees represented a variety of backgrounds; complete lists of both nominees and selected committee members are provided in Appendix E. The numbers of committee, excluding MSDE staff, members selected for each committee is listed below:

<u>Committee</u>	<u>Number of Members</u>
Program Specifications	36
Test Administration Specifications	23
Test Specifications:	
English	28
Mathematics	20
Science	32
Social Studies	28

Each of the four Test Specifications committees met for four working days (March 10-11, April 14 or 28, and May 15). On the morning of the initial meeting in March, MSDE staff welcomed the participants and then ETS and College Board staff outlined the tasks facing the Test Development committees, including an overview of the development process and the tasks that were to be accomplished at the meeting. The participants also received a set of working assumptions (see Appendix C); these assumptions were drafted in concert with MSDE staff. Committees were told to recommend changes in the working assumptions, defined as a set of assumptions that provide a common understanding so that participants could begin the task, whenever they felt it appropriate. Each content group then met separately for the duration of the session. The groups began by considering sample test specifications for one or more of the tests which had been prepared in advance by ETS test development staff. After this initial exercise, the English, Mathematics, and Science groups decided to work in test groups for part of the work period. By the end of the second day, initial specifications had been prepared for all twelve content tests which linked the Core Learning Goals (including the Skills for Success) to the test blueprint. In addition, the groups began to address other topics for the specifications: a) multicultural concerns, b) content concerns related to special populations, c) documentation of tools or equipment that would be needed, and d) items with cost or implementation implications.

The draft test specifications were shared with the Program Specifications committee at its first meeting (April 10-11) and with the Test Administration Specifications committee at its first meeting (April 11). These groups also received an orientation session to help them understand the HSA design process, to outline their specific tasks and responsibilities, and to review and comment upon the groups' working assumptions (see Appendix C). Each group was asked to review the draft test specifications and to provide feedback to the content groups. In addition, the PSC was asked to provide specifications recommendations for special populations (special education students, limited English proficient students, gifted and talented students), to address issues that cross content areas, and to check the links to Skills for Success. The TASC was asked to provide specifications recommendations for all aspects of test administration, to resolve issues related to the Preparation Plus design model, and to provide advice on staff development issues. Additional details about the activities of and recommendations from these two committees are contained in a later section of this report.

At the April Test Specifications meetings, each content group considered the feedback from the PSC and TASC groups. A second draft of the test specifications was then prepared. At the May meeting, each content group developed item-level specifications for each test and made final revisions to the draft test specifications.

1.3 *Level of Item Specificity*

The Core Learning Goals (CLG) for four content areas and the Skills for Success outline the essential skills and knowledge that Maryland students should be expected to know. In each area, these CLGs contain: a) goal descriptions, which reflect general learning goals and concepts, b) expectations, which reflect information about the nature of the fundamental information or processes to be addressed, and c) indicators of learning, which reflect specific activities, content, or skills to be covered. The content test specifications for each test have been written to require that all goals and expectations are covered. However, each indicator may not be covered in each test form; due to the number and/or nature of the indicators in some content areas, a subset of the indicators may be covered in any one test form while all indicators will be covered over a period of years.

In addition to specifying the distribution of the CLGs, the test specifications indicate which content will be tested in a selected response, brief constructed response, and extended constructed response format. Furthermore, item-level specifications have been included which address further concerns about the items to be used such as the kinds of items, the kinds of materials to be used, the distribution of items, and the nature and number of visual materials to be included. Additional question type information (e.g., percentage of “compare” items) may be added to the item-level specifications at a later point after MSDE has considered what is appropriate for the HSA program and after the test developer has had an opportunity to suggest additions to the possible question types.

1.4 *Overview of Subsequent Topics*

In the aggregate, the Test Specifications committees made hundreds of recommendations about the structure of the HSA tests. CB/ETS has endorsed the vast majority of these recommendations, which are contained in the specific specifications documents. In the very few cases where CB/ETS does not endorse the committee recommendations, we have made a specific recommendation in the text below. These situations, though rare, have typically involved issues related to the technical quality of the test or to the operational aspects of test administration. In only one instance CB/ETS has recommended deferring a topic for Maryland Board of Education approval: the structure of the English Preparation Plus model as modified by the English Test Specifications committee (this topic is discussed in a later section of this report).

In the following sections, general information is provided for each of the four content areas. In each content area, topics common to all tests have been identified, item types that will be used in the test have been defined, and proposed subscores have been outlined. In addition, the link to the CLGs (content and Skills for Success) have been reviewed, and non-testable CLGs, if any, have been identified. If any content committee recommendations contradict the recommendations of

another committee (e.g., Program Specifications or Test Administration Specifications), these areas have been listed and the CB/ETS recommendation included.

2. Information Common to All Tests

2.1 *Testing time*

The length for each test is three hours (180 minutes). Of the 180 minutes, 15 minutes have been allocated for directions (both test and item type instructions), leaving 165 minutes of testing time.

2.2 *Fairness*

In accordance with the Maryland State Board of Education Regulations Title 13A, Subtitle 04, Chapter 05 (“Education That is Multicultural”), HSA tests will include multicultural materials (textual, visual, and/or auditory) to the extent appropriate for the content of each test. Such topics, sometimes referred to as “minority topics” for space-saving reasons, are intended to cover various aspects of test content including names and/or portrayals of people, tapes of speeches, reading passages reflecting contributions of people of color, etc.

In addition to developing materials that are multicultural, reviews for multicultural and fairness concerns must be conducted at both the item and test level. At the item level, these reviews must ensure that the materials reflect Maryland’s multicultural society and student population. The reviews must also reflect the Maryland Board of Education’s commitment to develop test materials that are free of racist, sexist, or otherwise potentially offensive language and images. To this end, materials which involve stereotyping, inflammatory or highly controversial topics, or inappropriate tone would be considered unacceptable. At the test level, these reviews must also consider the balance and overall impression of the test.

2.3 *Number of answer choices for selected-response items*

There will be four answer choices for the selected-response questions. Answer choices for test items must not include “all of the above,” “none of the above,” or Roman numerals (sometimes known as “k-type” items) because these features tend to negatively affect question performance.

2.4 *Organization of selected-response questions (blocks)*

Selected-response items will be administered in blocks so that students will not be required to switch frequently between the selected-response answer sheet and the constructed-response answer booklet. This means that students will not be required to answer one or two multiple choice questions, then answer one brief constructed

response, then answer two or three multiple-choice questions. Once they move into a booklet, they will remain in that booklet for a fairly long period of time.

2.5 *Visuals*

Test specifications must limit the number and location of visuals that are printed in half-tone or color because such visuals add to the logistical printing problems and to the printing costs.

2.6 *Inappropriate context*

Because of its association with differential performance by women, sports scenarios must not be used as a context for items in HSA examinations.

2.7 *Dictionaries*

Dictionaries must be available to students throughout any test that permits dictionaries. Students must be aware that any time they spend using dictionaries will impact the amount of time they have to complete the test.

2.8 *Contradictions between committee recommendations*

2.8.1 *Multicultural representation*

Test Specifications committees in all content areas recommended a specification that requires that 15% of the test materials (text, graphics, etc.) represent the multicultural nature of Maryland's citizens (a number based on the minority proportion of the Maryland population). The Program Specifications Committee recommended that the percentage of multicultural material be equivalent to the percentage of minority children in the Maryland public school system (a number greater than 30%). If the larger number is being considered, it will be important to determine whether there would be any adverse impact on the content to be covered in English or Social Studies tests. (For example, would this requirement affect the number of particular topics or authors to be covered in such a way that it would be difficult to meet test specifications?) A further discussion of this topic is contained in a later section, along with the CB/ETS recommendation.

2.8.2 *Access to dictionaries*

In the four science tests, the content groups recommended that dictionaries not be available because too many science concepts (e.g., biology genus and species information) are included as part of the definitions of words. This recommendation contradicts the Program Specifications Committee's suggestion that dictionaries be available for students during all tests. In order

to maintain the validity of the content in the science tests, CB/ETS recommends that dictionaries not be available during these tests.

3. English

Summary information about the English tests is presented below, and detailed proposed specifications for the English 1, English 2, and English 3 tests are contained in Appendix C.

3.1 Common information for all tests

The English Test Specifications group has specified that they wish to use a Preparation Plus model for the test and that the test should be given on two days. As described by the group, this model calls for a 60-minute preparation period to be given on the same day as the first part of the test. (A full rationale for the Preparation Plus model is provided in the introductory material for the proposed English test specifications in Appendix C.)

The current structure of the Preparation Plus model is one that the State Board of Education must note since it differs somewhat from the model described to the Board in January. In the previous description, the preparation period was integrated into the course instruction time. In the current model, the preparation period immediately precedes the testing time.

3.2 Defining the Item Types

The English tests will include SR questions that ask about literature and about language usage topics. A full description of the item types to be used is included in the proposed English test specifications in Appendix C. Each of the English tests will include 90 selected-response operational questions; this number is in accordance with the psychometric recommendations made by CB/ETS to achieve appropriate reliability of .90.

The English tests will include both BCRs and ECRs. Each test will include several BCR questions which will test literary or composition topics. Each English test will also include one ECR on a composition topic.

Illustrative examples of SR and CR item types for the English tests are included in Appendix D.

3.3 Area subscores

The following subscores were recommended for each English test:

- a. Goal 1
- b. Goal 2
- c. Goal 3
- d. Goal 4

3.4 Link to Core Learning Goals and identification of any non-testable Core Learning Goals

All goals and expectations will be covered in the appropriate English test. Some of the indicators are not testable in the HSA paper/pencil assessments: 1.1.4, 2.2.6, 2.3.2, 2.3.5, 3.1.1, 3.1.2, 3.1.7, 3.2.1, 4.2.3, 4.3.2, and 4.3.3. Most of these indicators require students to present or judge oral presentations or to revise their own written compositions.

3.5 Link to Skills for Success

On each test, the links to the Skills for Success have been indicated for each expectation in each CLG. The links to Skills for Success are shown in Figures 1, 2, and 3 below.

Figure 1.

ENGLISH 1	SKILLS FOR SUCCESS EXPECTATIONS																		
EXPECTATIONS	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																			
1.2																			
1.3																			
2.1																			
2.2																			
2.3																			
3.1																			
3.2																			
3.3																			
4.1																			
4.2																			
4.3																			

NOTE: Expectations 1.4, 1.5, and 5.1 - 5.4 are not measurable.

Figure 2.

ENGLISH 2	SKILLS FOR SUCCESS EXPECTATIONS																		
EXPECTATIONS	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																			
1.2																			
1.3																			
2.1																			
2.2																			
2.3																			
3.1																			
3.2																			
3.3																			
4.1																			
4.2																			
4.3																			

NOTE: Expectations 1.4 and 5.1 - 5.4 are not measurable.

Figure 3.

ENGLISH 3	SKILLS FOR SUCCESS EXPECTATIONS																			
EXPECTATIONS	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	3.1	3.2	3.3	3.4	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																				
1.2																				
1.3																				
2.1																				
2.2																				
2.3																				
3.1																				
3.2																				
3.3																				
4.1																				
4.2																				
4.3																				

NOTE: Expectations 5.1 - 5.4 are not measurable.

3.6 Contradictions between content committee recommendations and other committee recommendations

The English Test Specifications committee has recommended that English tests be given on two days because it will take more than four hours to administer the preparation period and the full amount of testing time. The English Test Specifications committee is not averse to one day of testing, but they believe that students will be fatigued by four hours of test work. The Program Specifications Committee has recommended that tests be given on one day, and ETS also recommends that each English test be given on one day.

4. Mathematics

Summary information about the Mathematics tests is presented below. Detailed specifications for the Mathematics Test 1 and the Mathematics Test 2 are contained in Appendix C.

4.1 Common information for all tests

The Mathematics Test Specifications committee recommended that a graphing calculator with specified minimum capabilities be required for both mathematics assessments. They recommended that the calculator be one that students regularly use during instruction so that students will be able to use the calculator effectively.

The requirement of a graphing calculator has implications not only for equipment costs, but for staff development. Teachers must be able to use graphing calculators effectively in order to instruct their students. Given the current timeframe for administration of the no-fault tests, it would be difficult to provide that instruction in sufficient time for the initial no-fault administration. CB/ETS, therefore, recommend that the use of graphing calculators be phased into the Mathematics tests. This means that the initial year of no-fault administration would not require such calculators. During this year, staff development could occur so that calculators could be required in the second year of no-fault administration. (See additional discussion of this issue in a later section of this report.)

4.2 Defining the Item Types

One type of selected-response item (SR) will be used in both Mathematics tests: multiple-choice questions in which students select the single correct answer from among four choices. In addition, each test will include student-produced response (SPR) questions in which students are asked to generate a correct answer and grid this answer onto an answer sheet; no answer choices are provided for SPR questions. The Mathematics tests will contain 70 questions of the SR and SPR format unless no-fault data demonstrate that a greater number is needed to obtain sufficient reliability for high-stakes decisions. Some items will be discrete items and some will be linked in sets based on a common stimulus.

Two types of constructed-response (CR) items will be used in both Mathematics tests: a) brief constructed-response (BCR) questions in which students may be asked to provide a numerical and/or brief narrative response, and b) extended constructed-response (ECR) questions which will require explanations and may have multiple parts.

Illustrative examples of SR and CR items types for the Mathematics tests are provided in Appendix D.

4.3 *Area subscores*

The Mathematics content committee recommends that there be four subscores for Mathematics test 1:

- a. Goal 1, Expectation 1.1
- b. Goal 1, Expectation 1.2
- c. Goal 3, Expectation 3.1
- d. Goal 3, Expectation 3.2

The recommendation is for three subscores for Mathematics test 2:

- a. Goal 2, Expectation 2.1
- b. Goal 2, Expectation 2.2
- c. Goal 2, Expectation 2.3

4.4 *Link to Core Learning Goals and identification of any non-testable Core Learning Goals*

In both tests, the item specifications link directly to the CLGs for Mathematics. In the Mathematics test 1, all expectations in Goals 1 and 3 will be tested. In Mathematics test 2, all expectations and indicators in Goal 2 will be tested.

4.5 *Link to Skills for Success*

In both Mathematics tests, the links to Skills for Success have been made clear at the expectation level and, where appropriate, at the indicator level. Links have been made to many parts of Skills for Success Goals 2, 3, and 4, as shown in Figure 4 below.

Figure 4.

MATHEMATICS	SKILLS FOR SUCCESS EXPECTATIONS																	
EXPECTATIONS	1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	4.4	5.1	5.2	5.3	5.4
1.1																		
1.2																		
2.1																		
2.2																		
2.3																		
3.1																		
3.2																		

4.6 *Contradictions between content committee recommendations and other committee recommendations*

The Mathematics Test Specifications committee has recommended that the Mathematics tests be taken in two parts to be given on two separate days because they feel that answering Mathematics items is very fatiguing. This recommendation contradicts the Program Specifications Committee, which has recommended that all tests be administered on one day. CB/ETS recommends that the Mathematics test be given on one day because the administrative problems and costs rise when tests are given on more than one day.

5. Science

Summary information about the science tests is given below. Detailed specifications for the following tests are contained in Appendix C: Earth and Space Science, Biology, Chemistry, and Physics. These detailed specifications contain specific information about the equipment that would be required for the content to be tested.

5.1 *Common information for all tests*

The science groups recommended that laboratory work be considered a pre-requisite to all HSA science tests. This laboratory work is part of the focus of the Science Core Learning Goal 1, which delineates the skills and processes that students must develop in science. The Test Specifications groups did not specify particular experiments to be conducted, and the science tests do not assume that particular experiments have been done. However, the groups wanted to ensure that students have experience in conducting experiments so that they can become familiar with laboratory concepts and processes (e.g., formulating hypotheses, designing

experiments, collecting data), and have provided a list of suggested activities for each discipline that can be used in the classroom to address concepts in the Core Learning Goals.

5.2 Defining the Item Types

Each of the science tests will contain four types of SR items: discrete items (which can be individual items or items in a problem set which share a common stimulus), classification-set items (items in a set which refer to the same set of answer choices and which may also have some common stimulus material), laboratory-set items (which have a common stimulus material that refers to a laboratory situation and in which the items may assess laboratory skills), and technical passage sets (items in a set involving interpreting and analyzing a technical passage). Each of the science tests will include 90 selected-response operational questions; this number is in accordance with the psychometric recommendations made by CB\ETS to achieve appropriate reliability of .90.

Each of the science tests will include two types of CR questions: BCRs and ECRs. The BCRs will ask students to provide brief information about a topic; these topics may be based on a laboratory experience. The ECRs will require more time than BCRs and will ask students to generate a more complex response such as constructing a graph from data and then analyzing it or explaining a concept from a laboratory experience.

Illustrative examples of the item types for each of the science tests are provided in Appendix D.

5.3 Area subscores

The following subscores were recommended for each science test:

Earth/Space Science

- a. Goal 2, Expectations 1 and 2
- b. Goal 2, Expectations 3 and 4
- c. Goal 2, Expectations 5, 6, and 7

Biology

- a. Goal 3, Expectations 1 and 2
- b. Goal 3, Expectations 3 and 4
- c. Goal 3, Expectations 5 and 6

Chemistry

- a. Goal 4, Expectations 1 and 2
- b. Goal 4, Expectation 3
- c. Goal 4, Expectation 4

Physics

- a. Goal 5, Expectation 1
- b. Goal 5, Expectations 2 and 3
- c. Goal 5, Expectations 4 and 5

5.4 Link to Core Learning Goals and identification of any non-testable Core Learning Goals

In each test, the relevant Core Learning Goals and expectations will be tested in every test form except as noted below. In CLG 1 (which is contained in each science test), indicators 1, 2, and 3 of Expectation 3 are also not testable. In Earth and Space Science, expectation 8 overlaps Goal 1, and the indicators that are the same will not be tested twice. In Chemistry, test items will not be developed for expectations 5 and 6 because those expectations already appear in Core Learning Goal 1. In Physics, expectations 6 and 7 are not testable in the HSA paper/pencil assessment. These expectations require the student to investigate the impact of physics on society and to show linkages between physics and other areas of knowledge including language arts, fine arts, social studies, etc.

5.5 Link to Skills for Success

The links between Science tests and the Skills for Success occur in Science CLG 1. Links have been made to Skills for Success Goals 1, 2, 3, 4, and 5, as shown in Figure 5 below.

Figure 5.

SCIENCE EXPECTATIONS	SKILLS FOR SUCCESS EXPECTATIONS																
	1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																	
1.2																	
1.3																	
1.4																	
1.5																	
1.6																	
1.7																	

6. Social Studies

Summary information about the Social Studies tests are given below. Detailed test specifications for the Government, U.S. History, and World History tests are contained in Appendix C.

The Social Studies committee has recommended that each Social Studies test contain only 60 SR items. Based on past experience with tests such as the Advanced Placement examination (discussed further in Appendix B), ETS/CB believes that this number of SR items will not yield sufficient reliability for high-stakes decisions. We suggest that contingency plans such as testing a 90-item SR version be incorporated into the no-fault administration in order to help make the final decision about the number of SR items that is needed in social studies tests.

6.1 *Common information for all tests*

The Social Studies Test Specifications committee has recommended that a packet of materials be made available to teachers in advance of testing. This packet would provide suggestions about how to teach key skills, model instructional activities for teachers to use, illustrative examples of item types, and scoring rubrics. CB/ETS recommends that this type of packet be developed not only for Social Studies teachers, but also for teachers of each of the tests in the HSA program.

6.2 *Defining the Item Types*

Each of the social studies tests will contain 60 SR questions unless no-fault data demonstrate that a greater number is needed to obtain sufficient reliability for high-stakes decisions. Some of the SR items will be discrete items and some will be linked in sets based on a common stimulus.

The CR items on the social studies tests will contain both BCR and ECR questions. BCRs will allow for a brief development of an idea in social studies. Each test will also contain one ECR of approximately 30 minutes which will require an extended written response; the ECR may be a document-based exercise.

Illustrative examples of SR and CR item types for each of the social studies tests are included in Appendix D.

6.3 *Area subscores*

Recommended subscores for the Government test are:

- a. Goal 1
- b. Goal 2
- c. Goals 3 and 4

Goals 3 and 4 have been combined for the Government test because there are too few questions in each goal to provide useful information if the goals are reported separately.

Recommended subscores for the U.S. History and World History tests are:

- a. Goal 1
- b. Goal 2
- c. Goal 3
- d. Goal 4

6.4 *Link to Core Learning Goals and identification of any non-testable Core Learning Goals*

All Core Learning goals, expectations, and indicators will be covered in the social studies tests.

6.5 *Link to Skills for Success*

On each test, the links to the Skills for Success have been indicated for each expectation in each CLG. There are links to Skills for Success Goals 2, 3 and 4, as shown in Figures 6, 7, and 8 below.

Figure 6.

GOVERNMENT	SKILLS FOR SUCCESS EXPECTATIONS																		
EXPECTATIONS	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																			
1.2																			
2.1																			
2.2																			
3.1																			
3.2																			
4.1																			

NOTE: Expectations 1.1 - 1.5, 3.3, and 5.4 are not measurable.

Figure 7.

U. S. HISTORY	SKILLS FOR SUCCESS EXPECTATIONS																		
EXPECTATIONS	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																			
1.2																			
2.1																			
2.2																			
3.1																			
3.2																			
4.1																			

NOTE: Expectations 1.1 - 1.5, 3.3, and 5.4 are not measurable.

Figure 8.

WORLD HISTORY	SKILLS FOR SUCCESS EXPECTATIONS																		
EXPECTATIONS	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																			
2.1																			
2.2																			
2.3																			
3.1																			
4.1																			

NOTE: Expectations 1.1 - 1.5, 3.3, and 5.4 are not measurable.

IV. TECHNICAL SPECIFICATIONS

The technical specifications for the Maryland HSA tests (contained in Appendix B) are designed to ensure that the tests meet all of the psychometric standards endorsed by national educational organizations such as the American Educational Research Association and the National Council on Measurement in Education. This section will provide a brief overview of the detailed specifications.

One of the major concerns in the design of any assessment is the validity of the measure. The HSA technical specifications address the ways in which the validity of the tests can be accomplished: through the foundation of the core learning goals, through the consistency of curriculum and instructional practices across educational districts, and through the appropriate use to be made of the tests. These approaches can help to ensure that the content validity of the tests can be demonstrated, and that the validity of decisions arising from test scores can be established.

Another major concern in test design concerns the reliability of the tests. The technical specifications provide information to help ensure that the tests reliably measure what they are designed to measure. Because different types of test questions affect reliability in different ways, the technical specifications have addressed the need for appropriate numbers of selected-response and constructed-response questions in each of the four content areas. These specifications have also noted that reliability for multiple decisions (such as remediation or high achievement) can affect the ways that tests are assembled. Therefore, a priority should be established for the several proposed uses of the HSA tests so that tests can be designed to support the most essential use(s).

In order for the HSA tests to move smoothly into an operational phase, they must be designed to address issues related to field testing, calibrating, and scaling test questions. The technical specifications contain recommendations about the ways that these activities can be conducted for both selected-response and constructed-response test questions. They also contain discussions of several equating procedures that could be used.

Since the HSA tests will contain several different question types, which may be scored in different ways, the technical specifications contain recommendations about scoring issues for both machine-scored and hand-scored test questions. For the machine-scored questions, these recommendations concern test questions to which examinees do not respond. For the hand-scored questions, these recommendations concern procedures intended to maximize scorer reliability.

The topics of test and item analyses have also been addressed in the technical specifications. Recommendations have been made for appropriate analyses of individual test questions, as well as for whole tests. The issue of weighting of the various types of items (selected response and constructed response) has also been discussed and a recommendation has been made about what to consider in making weighting decisions.

Finally, the potential reuse of test questions has been discussed and recommendations have been provided for both selected-response and constructed-response questions. The ease with which the constructed-response questions can be remembered has led to a recommendation that they be used far less frequently than selected-response questions.

V. ASSESSMENT PROGRAM SPECIFICATIONS

This chapter addresses major program specifications for the High School Assessment (HSA) such as security, scheduling tests, equipment required for the tests, and other factors that address the operational aspects of the assessments. CB/ETS worked collaboratively with four content-based specification teams in developing proposed specifications for each of the 12 tests described in chapter 2. CB/ETS also worked with two committees designated by MSDE to address these administrative issues—the Test Administration Specifications Committee (TASC) and the Program Specifications Committee (PSC). In some instances, the test specifications committee for a specific content area (i.e., English, math, science, social studies) proposed procedures or specifications that were not supported by the TASC. In other instances, the TASC proposed administrative procedures that were not supported by the PSC. Because no one group had ultimate authority for any decisions, a large proportion of design specifications remain undecided at this time. Until some group is authorized to make decisions regarding these specifications, many critical design features of the tests cannot be determined, and test development cannot be finalized. If test development proceeds before these decisions are made, it is highly likely that the design and development will require substantial modifications at a latter stage, which will result in additional costs, scheduling delays, and other inefficiencies.

A number of disparate recommendations emerged between groups on issues central to the HSA. CB/ETS, along with staff from MSDE, did not have adequate time to attempt to resolve these disagreements among the groups. Rather, our charge was to: (1) report the recommendations formulated by each committee; (2) evaluate the feasibility and potential consequences of these recommendations on the HSA; and (3) where necessary, provide additional recommendations that best ensure a technically acceptable, administratively feasible, educationally sound, and economically viable alternative for central HSA issues. When recommendations from committees did appear feasible, we did not provide a separate recommendation from CB/ETS.

There are a number of issues that these committees did not have adequate time to consider and they were deferred until future meetings in the fall. Because work conducted by CB/ETS will end in August 1997, we have chosen to briefly address each issue at this time and provide some broad perspectives to MSDE on potential solutions to all issues, even those that have not yet been considered by the TASC and PSC. To delay some of these issues until late fall would endanger the current timetable for field testing and implementation of the HSA.

Four separate test specification committees were established by MSDE with over 100 Maryland educators participating. Committees were established for English, mathematics, science, and social studies. MSDE used a comprehensive process to solicit nominations from all Maryland school districts, and final appointments were made in February to ensure each committee was representative of the diversity of districts, students, and domains within each subject area. Appendix E contains a roster of all

committee members who were appointed. The test specifications committees met on three occasions for a total of four days. They met as a single committee according to subject area, but also spent some of their time meeting as separate subgroups to consider each HSA test within a content area (e.g., the science committee divided itself into biology, chemistry, earth/space science, and physics). Any recommendations from a test administration committee that departed from the standard design, administrative conditions, or scoring assumptions (assumptions developed to ensure some level of consistency in the format, appearance, administration and scoring of all tests) were forwarded to the TASC for review.

The TASC and PSC met on two occasions. Approximately 60 additional Maryland educators were appointed to these committees using the same nomination process described above. Appendix E contains a roster of all committee members who were appointed as well as a summary of the meetings. All meetings of phase 2 activities occurred within a 60-day period of time (from mid-March to mid-May).

1. Security

Generally, the Test Administration Specifications Committee (TASC) recommended that MSDE develop few specifications concerning security issues; rather, they felt that schools should have maximum flexibility in determining how to best ensure security and minimize risk of cheating. They recommended that MSDE suggest best practices, while permitting schools to address the requirement of test security, and that school officials be held accountable for any security lapses. They further suggested principals, local accountability coordinators, and test administrators sign a form attesting that they had implemented appropriate security procedures to ensure the validity of test administration.

High-stakes testing programs require a level of standardized administrative conditions that argue against substantial flexibility among schools in determining how to best implement procedures. Rather, detailed security specifications should be developed by MSDE, although some level of flexibility will certainly be required on several issues. Schools that may be unable to conform to one or more of the specifications would then need to contact their local accountability coordinator, who would seek an exemption to the specification. Test Administrator Manuals have been developed for the Maryland Functional Tests, the Maryland Writing Test, and the Maryland School Performance Assessment Program. A similar administrator manual should be developed for the HSA. MSDE should review procedures in place in these testing programs, as well as the accompanying documentation, and develop specifications for the HSA prior to the first no-fault administration.

In reviewing the test administrator manuals for these testing programs, many of the procedures used for existing Maryland testing programs may be successfully applied to

the HSA. However, there are some areas where changes appear needed and some areas of program policy which may require further examination. Specifically:

- Distribution and Collection of Materials - HSA will require a substantially greater level of effort and coordination from school test coordinators because one or two specific assessments are envisioned to be administered each day for a number of consecutive days. Materials will need to be distributed to test administrators prior to the date of each test, and materials will need to be collected separately for each assessment on each day of testing and bundled and shipped separately. Once the HSA is fully operational (with 12 separate tests), these processes may require five to seven times the level of effort associated with the Functional Tests.
- Tests should not continue across two days in a high-stakes program. If a test requires two days for administration, this would add one day of testing per semester (and may translate to two additional days without instruction under some models discussed later). It would also introduce additional problems if students are absent for any one of the two days and create formidable challenges for schools who will be required to match testing forms.² The Program Specifications Committee recommends all tests be completed within one day, and CB/ETS believe that this is an appropriate resolution.
- The Functional Testing Program and the Maryland Writing Test each permit students as much time as required to complete the tests -- referred to as "unspeeeded testing." Students who have not completed the former tests within a specified period of time are often required to change locations to complete the testing. These procedures (and the practice of extending administration of any test beyond one day) do not appear to be optimal for the HSA for a number of reasons: (a) such variable time will introduce a source of invalidity to the tests (this is a substantial concern with assessments which are geared to high standards); (b) these procedures will create tremendous operational difficulties for schools if they are to administer two or more assessments on the same day (as CB/ETS believe is required for administrative feasibility because administering one test per day would extend testing for over two weeks each semester); (c) additional security concerns would be introduced; and (d) these procedures cannot be easily accommodated in a 3-hour test where students complete two separate portions of a test without returning to sections after they have been completed (as explained later) and will introduce additional nonstandardization to the HSA. Rather, a model similar to that employed with MSPAP (having standard testing time requirements) is appropriate with the HSA. As with other state tests, items not reached would not be scored as zeros. MSDE can still ensure the tests are unspeeeded (or minimally speeeded) by adjusting the final test specifications to ensure a very high level of completion.³ The Program Specifications Committee recommends all tests have the same time limit.

² While these procedures have been employed in the Writing Test, the volume of the HSA may be four or five times larger and introduce substantial challenges and costs.

³ Most high-stakes testing programs use criteria to ensure tests are not highly speeeded. In the case of the

1.1 *Student Identification*

What precautions will be in place to prevent individuals from completing the HSA for fellow students (e.g., fellow students completing a test for a friend and using his or her name and identification information)? In all high-stakes testing programs, a few common methods of cheating have been reported. One form of cheating is to recruit a confederate (a fellow student or confederate) to pose as the student, sign the false name, and complete the test in his or her place. Many high-stakes testing programs require a picture identification from students to verify their identity prior to administration of the test.

The Test Administration Specifications Committee (TASC) has recommended that student identification *not* be required of students; many schools in the state do not issue such identifications to students. The TASC has recommended that school staff sign a form verifying that student names and identification information on each test form are accurate (the student completing the HSA is in fact the student whose name and identification information is on the test) and that each school or district develop its own procedures.

Requiring students to produce a picture identification is probably the most successful mechanism to prevent this form of cheating, especially since the TASC has recommended that schools be allowed flexibility to administer the HSA in classrooms or in large administrative settings (the same flexibility provided for the Maryland Functional Tests). Student identifications are not needed when school staff (usually the students' teacher) are administering the tests and can quickly determine that students completing the test are all in his or her class. However, if the classroom teacher is not the test administrator (or a proctor), as may be the case in large group administrative settings, there will be no guarantee that the assigned administrator will be able to recognize each and every student in that room.

Even with this concern, we believe that alternative procedures can be developed by MSDE which would also address this issue. However, CB/ETS must disagree with the second component of the TASC recommendation that would permit each school or district to develop its own procedures and have a school official sign to verify that security was upheld during testing. One alternative would be to have schools require the classroom teacher to serve as the test administrator or proctor for administrations of assessments to all of his or her classes (e.g., U.S. History,

SAT, two criteria are employed: (a) all students should complete 80% of the test form, and (b) 85% of students should complete the entire test. Plus, not reached items are not counted as zeroes. When higher levels of test completion are employed (e.g., 95% of students should complete the entire test) there is a risk that there will be an increased amount of time between the time the first students and the last students complete the test. This situation could cause additional anxiety for the last students still working on the test (who see their peers completing the test and leaving the room), as well as introduce substantial operational and security issues for the school (see section II, 9.3).

Biology, English 1). The teachers would then visually complete a seating chart for the class (or large-group administration) and sign a statement verifying that they visually recognized and knew the identity of each student in the setting. In any case, a seating chart (or a student roster at the minimum) should be completed and returned with the test materials.

1.2 Informational Materials For Parents

The TASC recommended that several pamphlets be developed by each district, based on state guidelines and policies, which would outline appropriate and inappropriate student conduct in preparation for and completion of the tests. This brochure would then be distributed to each student and his or her parents. Students would then sign a form attesting that they had read and understood the policies and consequences of inappropriate testing behavior prior to taking any tests.

Such a procedure would be very helpful in ensuring that students and their parents are informed, in advance, of proper and improper testing behavior. Students' signatures would also provide documentation that they had received the information. However, the materials should be developed by MSDE to ensure all students receive the same information in a consistent manner. Local development is unnecessary, would create additional and unnecessary work for districts, and result in students receiving slightly different information or messages.

The TASC and PSC also recommended that additional materials describing the purpose of the HSA be developed and shared with students, parents, teachers, staff, and the general public before students enter middle school. The PSC recommended sample items be included in some of these materials.

1.3 Scrambling Of Multiple Choice Items

The TASC agreed that scrambling of multiple choice items across two or more forms is desirable if it will increase security, reduce the need to create more burdensome security procedures (e.g., ensure students sit 3 feet apart during testing), and will not impact test performance. This procedure permits varying the order that MC items are presented across several test forms reducing the possibility that students sitting in close proximity can copy responses from the answer sheets.

1.4 Additional Procedures

Sample materials outlining several additional procedures for test security were distributed to the TASC for review. However, the TASC generally felt that the development of additional test security procedures for administration or post-administration analyses could be deferred. Basic security procedures and the types of post-administrative inquiries and analyses that will be implemented should be

developed this fall to permit test development to proceed on schedule. Sample materials used in College Board testing programs are reprinted in Appendix F.

2. Administration

As noted above, many features from the existing Test Administrator Manuals for other state assessments can be applied to the HSA. For example, the Code of Ethics, policies on snow days, responsibilities of key staff, prohibition on teaching to the test, and access to secure testing materials can be applied to the HSA with minor modifications (but in some cases more elaboration). MSDE should review procedures in place in these testing programs, as well as the accompanying documentation, and develop specifications for the HSA prior to the first no-fault administration.

In reviewing the test administrator manuals for these testing programs, many of the procedures used for existing state testing programs may be successfully applied to the HSA.

2.1 Testing Time

The HSA will be comprised of 12 assessments, each of which is assumed to require 3 hours of testing. Of the three hours, 165 minutes will be devoted to the testing time and 15 minutes will be used for instructions. However, a portion of the 165 minutes will be used to administer pretest and/or equating items (as described in chapter II, 5). The TASC recommended that there be one or two breaks, depending on the preference of individual districts or schools, and the duration of breaks should vary depending on school logistics. Some members felt that individual schools should determine these issues. There was some concern that schools may have limited rest room facilities, and having a common break for all students completing one exam (in some cases this may be an entire class or 1/4 of the student population) would be impractical. The distance to the rest rooms and other factors were a source of concern. The PSC was divided between whether no break or one break was optimal.

The number of breaks provided to students may appear to be a trivial issue, but significant problems and additional costs will be introduced if a fixed number of breaks are not pre-determined and schools have flexibility in determining this issue. Students can only be given a break, permitting them to leave the testing room and use school facilities, as well as talk with fellow students, at the end of a test section. When returning to the test, students should not be permitted to return to sections of the test completed prior to the break (for security reasons). It would be quite easy for students to exchange answers and thoughts during the break if they were permitted to return to the same tasks exposed to prior to the break. If variable breaks are considered, MSDE may need to print test booklets in a way to permit schools to have either three 60-minute sections or two 90-minute sections. While a test can be physically printed to accommodate either structure of sections (2 or 3),

substantial costs would be involved in printing each test form under both arrangements to permit flexibility.

One break would appear adequate for students in 8th-12th grade. Schools may be permitted flexibility in determining the length of the break as long as test administration doesn't extend beyond one-school day. Generally, breaks of 15-20 minutes should be adequate. If a particular school does have severely limited rest room facilities, that school may use a staggered schedule (1/3 of the classes begin testing at 8 a.m., 1/3 begin at 8:30 a.m., 1/3 begin at 9 a.m.), which would result in staggered breaks for different classrooms. A potential schedule for a single test may be:⁴

7:30 a.m. School day starts
7:40 a.m. Students report to assigned classrooms for testing
7:50 a.m. Instructions (15 minutes)
8:05 a.m. Testing (section 1) (75 minutes)
9:20 a.m. Break (20 minutes)
9:40 a.m. Testing resumes (section 2) (90 minutes)
11:10 a.m. Testing Completed

If during testing time a student needs to use the rest room, he or she should be allowed to leave the test and then either: (a) return to the test with no additional time, or (b) have an invalid test and re-take the test during the make-up period. Section 4 raises related issues that should be considered in the overall context of scheduling the HSA test administration.

2.2 *Unspeeded Tests*

This issue concerns whether all students should be required to complete the test within three hours or have as long as needed to complete the test. If students are permitted to have as long as they require, the test is no longer a three-hour test (even though it may be developed so "most" students can complete it in three hours) it becomes a variable time test. As noted, the TASC has recommended that tests be unspeeded (which is variable time) and that each student be provided with adequate time to complete each test. The Program Specifications Committee (PSC) has recommended that each test have the same time limit. The TASC recommendation for unspeeded testing would actually result in a variable length test which contradicts the HSA assumption that each test will be three hours in length. If a test is developed to be completed by most students in three hours, the test may actually require about 1.5 to 2 times as long for all students to complete it. Therefore, this recommendation would result in a test length of 4.5 hours for some students and would reduce instruction time. Unlike the Functional and Writing Tests, unlimited

⁴ An illustration of the time required to complete one HSA test in the a.m. Part 4 in this chapter discusses alternative models where two or more HSA tests would be required.

time would be unfeasible for the HSA because it involves 12 tests across 4 grade levels and is developed to be a high standard assessment program. CB/ETS recommend a standard time be adopted for all students and that tests have a minimal level of speededness. Of course, this discussion does not apply to accommodations that will be provided to individual students based on a disabling condition. In those instances, individual accommodations—such as extended time—will be required and provided during individual administration of each HSA test.

2.3 Test Grouping

The TASC recommended that schools be permitted to administer the test in groups and locations (classrooms, auditoriums, cafeteria, lab settings, etc.) that are most practical. Some potential groupings for test administration include:

- Intact Classroom administration - students in a math class would report to an assigned room and complete the test in a classroom with a proctor.
- Random Classroom administration - students in 10 math classes would be assigned to 10 different classrooms in some random order (intact classes would not be tested).
- Large Group administration - two or more classes would be combined and administered the test in a large group setting.

Since multiple formats are presently used in the Functional and Writing Tests (although not with MSPAP), this recommendation may be feasible for the HSA. However, a number of issues must be examined such as: (a) the physical facilities of potential testing environments, and (b) real and perceived effect such variable test administration may have on the actual and perceived security and comparability. MSDE should not underestimate the public's concern with issues on comparable and equitable treatment of students. Differences in test setting may not be a problem with minimum competency tests, but such seemingly minor issues may become a concern when a large proportion of students do not pass an exam. Further study is required in the next year.

2.4 Test Administrators and Proctors

Eligibility to serve as test administrators and proctors, as defined in the MSPAP Test Administrator and Coordination Manual, appears adequate for the HSA. Of course, all test administrators and proctors will need to complete training prior to the administration of the HSA. The TASC recommended that Local Assessment Coordinators (LACs) determine the appropriate procedures for training test personnel.

This recommendation is problematic. MSDE should develop standard training materials, train all LACs prior to the first state-wide field testing, and monitor and evaluate training of test administrators and proctors each year. The availability of

standard training materials from MSDE should also aid LACs and be less of a burden to schools that would not need to develop their own procedures and materials. Because the HSA will be a high-stakes individual assessment program, MSDE needs to establish procedures (e.g., checklists which are completed and initialed by LACs and test administrators for each administration) to ensure they have adequately previewed all materials prior to test administration; understand how to securely store, bundle and return materials; are aware of the specific tools students may use and may not use during testing (and have arranged for approved materials to be present during testing); are aware of approved and prohibited instructional aides that may be displayed in testing environments; and understand procedures and policies that address the standardization, comparable administrative conditions, and security of test materials and test scores.

2.5 Test Administrations

MSBE has approved two annual test administrations with a make-up test date for each administration. The need for additional administrations may be determined during the state-wide field testing. The TASC has supported this recommendation, but noted that MSDE will need to justify this to districts which operate on the quarter system or offer half-credit courses. The PSC has recommended the Board reconsider this recommendation and add a third administration for the summer.

3 Student Populations and Test Taking

3.1 Completing Two Tests from the Same Content Area in One Test Administration Period

Should students be permitted to take two tests from the same content area (e.g., English 1, English 2) during the same test administration period? The TASC recommended that students be permitted to take two tests during the same administration time frame (e.g., spring 1999) if they were re-testing on one of the tests they had failed earlier. In this case, a student might re-take the English 1 test they previously failed and also take the English 2 test for the course they are currently enrolled in.

3.2 Taking a Test Without Taking the Course

Should students be able to take a test when they have not taken the corresponding course? The PSC has recommended that districts determine this issue; however, such inequity could raise basic concerns for the state assessment program. AP does allow students to take a test without enrolling in the course, but this policy is uniform across all schools and states. Students would still be required to complete most of the courses even if they passed the corresponding HSA based on current course requirements for graduation. If students are permitted to complete (and “pass out of”) a test prior to completing the respective course, a number of

criticisms are likely to emerge. For example, critics will question the rigor of a test that students can pass without completing instruction, or the rigor of a course that students do not need to complete for successful performance on the end-of-course test. Local and state requirements may still mandate completion of the course even after students pass the HSA test.

3.3 Students Not Enrolled in Maryland Public Schools

Should students attending private schools be permitted to take the HSA exams? Should students schooled at home be permitted or required to take the HSA exams? The TASC recommended that only those students enrolled in Maryland public schools be permitted to participate in the program—alternative procedures should be developed if MSBE wishes to have home-schooled students or private school students participate in the program. No formal recommendation came from the PSC, but several members felt that the HSA should be required of home-schooled and hospital-schooled students if they are seeking a Maryland diploma. In addition, some PSC members stipulated that out-of-state students be permitted to complete the HSA if it will be associated with merit or preferential admissions. The feasibility of testing out-of-state students and students from private schools presents substantial problems and may need to be deferred.

3.4 Transfer Students

A number of complex issues are involved in determining how and when to include transfer students in the HSA program. The PSC recommended:

- Students transferring from out-of-state or private schools be required to complete and pass any HSA that corresponds to a course they must complete in Maryland public schools. However, there was no agreement about what proportion of a course a student must complete to be required to complete the test. Some participants felt that a student should be exempt from the exam if he or she transfers into a Maryland school at a time when less than half of the course in Maryland. Others believe individual schools should determine this policy. CB/ETS suggest a standard policy be considered that would require transfer students to complete the HSA test if they have transferred in a course for 75% or more of instructional days. If students enroll in courses after more than 25% of instructional days have elapsed, students would not be required to complete the HSA tests corresponding to those courses.
- Students transferring within Maryland districts must complete and pass any tests that they have not already passed. Additional issues needing resolution were identified (e.g., students transferring from semester courses to full year courses, students transferring from integrated math to algebra).

3.5 *Advanced Placement and International Baccalaureate*

The PSC recommended that no exemptions be permitted for either AP or the IB. Their rationale was that most AP courses would be taken in 12th grade and that IB does not have adequate standardization to substitute for Maryland courses. This recommendation would likely reduce AP enrollment across the state over time. For example, students faced with the requirement to pass a HSA U.S. History or Biology test to graduate may opt out of AP courses to ensure they are prepared for the required tests offered by the state. Few students would likely then re-take the same course simply to attain college credit (in some instances local policy may prohibit students from retaking the same course). Such an outcome should be closely examined since it could substantially reduce Maryland student participation in these honors programs and be perceived as penalizing the highest achieving students. Currently, Maryland ranks sixth in the nation in AP exam participation with 188 per 1,000 students completing AP examination. Concordance or predictive studies could be conducted to determine if AP courses are of the same rigor or greater rigor than corresponding standard courses.

3.6 *Accommodations for Students with Disabilities*

The TASC briefly discussed this issue and determined that no action was required at this time and deferred further action until the existing State Advisory Committee could be formed in the summer to review the issue. The PSC arrived at several recommendations:

- Accommodations used for the test must be the same as those used for instruction.
- Students must receive the accommodation in instruction and be familiar with the accommodation prior to testing.
- A meeting of special education directors must be convened to resolve a number of issues concerning: (a) access to all courses, (b) effect of tests on drop-out rates, (c) staffing shortages for individual administration of HSA, and (d) whether the English tests be administered orally and still be a valid measure of the construct
- Braille forms should be developed where feasible. If material cannot be Brailled, it should not be scored. At least 80% of a test form should be able to be transferred to a Braille form.
- Student records should document the accommodations provided, but scores should not be flagged.

Generally the types of accommodations and policies need to be seriously examined in the near future in order to permit test development to proceed on schedule. MSDE's *Requirements for Accommodating, Excusing, and Exempting students in Maryland Assessment Programs* (9/17/96) already provides specifications for accommodations required for scheduling, settings, equipment, presentation, and response in each of the assessment programs.

The list of available accommodations appears appropriate for the HSA, but the final specifications must be developed. Setting and scheduling accommodations may be most easily applied to a new testing program. Under equipment accommodations, Braille test forms may present the biggest obstacle as some stimulus materials developed by the test specifications committee may not lend themselves to a Braille form (e.g., a political cartoon). If a substantial portion of any test cannot be converted to Braille, it would be inappropriate to produce a Braille form (since the content coverage and construct measured by the Braille form would not be equivalent to the test). The Test Developer will need to review the specifications with representatives from special education to first determine the percentage of tasks which can be converted to a Braille form. Test specifications may need to be modified if producing a Braille form of each test is a priority. If Braille forms cannot be produced, exemptions for visually disabled students would be granted. A minimum of 80% of a test should be capable of being translated to Braille, but higher proportions of the test items would be ideally sought. Oral presentation of tasks and cassette forms of the tests may present similar problems, but to a lesser extent. Large print forms should be produced for each test. Computer delivery of the tests was recommended by some groups during public engagement. This accommodation is not currently provided for other tests (although word processors are permitted) and would present significant technical issues and substantial additional costs. Extended time and individual administration of tests are common accommodations for most testing programs and should be easily incorporated into HSA (also see chapter IV, 8).

3.7 Limited English Proficient Students

Students with limited English proficiency are able to request, with approval from a multidisciplinary team, a one-time exemption from current state assessment programs, and similar procedures should be considered for the HSA. The PSC recommended that scorers for constructed response tasks receive additional advice to help them accurately score responses from these students and indicated that special scorers may be needed. In addition, they recommended a special meeting of LEP directors be convened to address other issues.

As with accommodations for students with special education, these issues should not be deferred further. Draft accommodations should be developed this fall, as well as a thorough review of draft test specifications to determine the extent they will create barriers to common accommodations already provided by Maryland. Developing alternate forms in other languages would present a major financial and psychometric challenge for MSDE and all committees agreed tests should be produced in English only (in addition to Braille and other accommodations for students with disabilities).

4. Scheduling

4.1 *Tests Administered at Same Date Across State*

Because only one form of each test will be initially developed for each administration, a state-wide test administration schedule specifying which tests are administered on each day will be needed. That is, schools cannot have flexibility in determining which day within a two-week testing period they prefer to offer each test. If the same tests are not administered on the same day across all districts, then students in districts testing later in the testing window will be able to determine the content of the tests from students who test earlier in the process. This is a particular problem for extended constructed response items where a disproportionately high proportion of the test score may be based on one or two extended tasks. If students receive advance knowledge about the topic and questions, they would have time to research and prepare a more comprehensive response. Students might also receive assistance from others in such advance preparation, creating an inequity across districts. This policy was endorsed by the TASC; however, they recommended that schools be given latitude in determining the start time for the exam.

Because school schedules differ across districts and schools, there will certainly need to be flexibility in the time tests are administered across schools. One mechanism to permit a substantial amount of flexibility while retaining quality control and minimizing security risks is for MSDE to specify a “window of time” when testing must begin. This window of time should not exceed 90 minutes (this would prevent students on a break in one district from contacting students beginning later in the day at another district). For example, MSDE might specify that each test must begin between 7:15 a.m. and 8:45 a.m. for tests administered in the morning.

4.2 *Test Administration in 1 or 2 Days*

The English and Mathematics Test Specification Committees have recommended that each of their assessments be administered across two days. Mathematics recommended that 90 minutes of testing be conducted on each of two consecutive days. English recommended that the 60-minute Preparation Plus period and a portion of the 3-hour test be conducted on the first day, with the remainder of the test administered on the second day. The TASC met with members from the English Committee but rejected this recommendation and stated all HSA tests should be administered on one day. Members of the TASC noted that the recommendations from English and math for two-day testing would: (a) result in a much higher rate of incomplete tests (some students may be absent on any one day and invalidate the test); (b) add significant burdens to schools in order to coordinate this aspect of the testing (e.g., pass out and collect the same materials in the same order); (c) require additional days for testing and decrease school days used for instruction (2.5 to 4 additional days of testing would be added, and potentially 2.5

to 4 less instructional days would be contained in a school schedule); and (d) raise security issues for the program. In addition, the cost of test administration would increase significantly across the state for each additional testing day required (e.g., substitutes, test administrators and proctors, schools days for non-instructional activities will increase the school calendar in some districts). MSDE will need to determine if the benefits of extending testing over two days in English and math would outweigh these serious challenges to administrative feasibility, cost effectiveness, validity, and security that have been identified by the TASC. If testing in any subject area is extended beyond one day, the discussion and schedules provided below will need to be revised. CB/ETS recommend that each test be completed in one day for the initial operation of the HSA. MSDE can later consider modifications once the initial program is up and running successfully.

4.3 Instruction During HSA Testing

This is a complex administrative issue concerning whether schools will be open for instruction on days when the HSA tests are administered. The alternatives for this issue depend to a large extent on MSBE's decisions concerning the scheduling of HSA tests (as discussed in item 2). For example, if MSDE determines that only one HSA test should be offered on a given day, that would require a minimum of 13 days of testing each year for schools on a full year schedule (and 26 days of testing each year for schools on a semester schedule). It may be untenable for schools to close for instruction an additional 13 or 26 days per year. Instead, schools may need to develop a complex instructional schedule for testing dates that would permit students not completing the test (e.g., English 1 on Monday) to attend their classes.

An illustration of just some of the problems with this alternative emerge when you consider an actual example, such as administration of English 1 on Monday from 8-11 a.m. In this example, classes may return to the normal schedule in schools after the English 1 test is administered (after the fourth period), but scheduling for the first four periods would require major modifications because teachers of English 1 (who may likely teach other courses during the first four periods) and classrooms (which are normally used by other classes during the four periods) will be unavailable for instruction. The scheduling problems with this scenario may be unresolvable and some schools may be required to either begin school later on testing dates or to not hold instruction at all. Of course, some schools may have adequate space for large group testing and the major issue would be obtaining adequate numbers of certified test administrators and proctors rather than scheduling of classrooms.

Another option is for schools to begin the instructional day after the test administration is completed. This would require students not completing the test (75% of the student body or more) to begin the school day around 10:30 - 11:00 am on 13 days (or 26 days for schools operating under a semester block schedule). A variation of this model is to have no instruction on these days which would be a

significant loss in instructional time. Both of these two alternatives may be equally unacceptable to some schools and parents. In addition, bus schedules and additional transportation needs may prohibit this option.

A third alternative that appears most feasible is to schedule two tests for each day -- this would result in all tests administered in seven days, or 14 days for schools on a semester block schedule (rather than 13 or 26 days respectively). First, MSDE would need to determine which tests are most likely to have non-overlapping student populations (minimal conflicts so few students would need to take two tests on any given day). For example, students taking the English 1 exam are not likely to be taking English 2 on the same day. Similar rules can be adopted for other tests (e.g., Algebra and Physics, US History and World History) although there will be instances when the conflicts will emerge for individuals students (e.g., students retaking a specific test) or districts with a less common curriculum sequence. Clearly, such conflicts will exist in *any* standard administrative schedule, but a make-up day could be used to handle these conflicts.

If MSDE determines that scheduling two HSA tests on the same day is optimal, then there are still at least two different methods for accomplishing this objective. Both tests could be offered at the same time (e.g., 8-11 a.m.), or one test could be offered in the morning session and another test offered in the afternoon session. This latter model would eliminate scheduling conflicts, but may require a small proportion of students to complete two exams on the same date (this method is employed with the Advance Placement program and is often used by colleges and universities for final exams). Schools may be better able to cancel instruction on these dates and only hold morning or afternoon testing sessions. Since few students would be required to take two tests on the same day, limited cafeteria service may be possible, but a modified bus schedule would be needed for the morning and afternoon testing sessions. Still, this may be the most feasible model given all the logistical problems with any of these alternatives.

The TASC did not identify a preferred model, but did recommend that each district determine whether or not instruction would be canceled on testing dates. There are a number of decisions regarding scheduling that are interrelated. Substantially more input from the TASC and PSC is required during the fall to develop optimal schedules. These alternatives are briefly summarized below.

4.3.1 Will students attend school for classes during HSA test administration?

Alternatives where all students attend classes and only students completing the HSA are exempt from the class.

Alternative One - Conduct regular classes for students during each period, but students completing the specific HSA test would go to assigned classrooms or

alternative locations (missing their first 3 or 4 periods of classes) to complete the test.

There are a number of obstacles in attempting to schedule instruction while HSA exams are administered. Schools wishing to pursue this will need to consider a few issues first:

1. Are there adequate testing locations with writing surfaces to accommodate up to 25% of the school's students and have all scheduled classes meeting during the instructional day? ⁵
2. Are there available qualified test administrators and proctors to conduct testing if all teachers are conducting instruction?
3. Given that students completing an HSA test will be absent from any courses scheduled during the corresponding time, will this create any insurmountable obstacles for students and staff (e.g., students completing English 1 in the morning cannot attend their first, second and third period classes that same day)?
4. Are there sufficient staff to handle the alternative scheduling arrangements for the HSA days?

For tests where a substantial portion of students are expected to complete the exam (15-25% of students), these issues are a concern. Because the volume for Physics and Chemistry tests may be very small, perhaps less than 5% of students, it may be inefficient to close school when so few students will take these tests.

Alternative Two - Begin the instructional day late (around 10:45 - 11 a.m.) after the HSA is administered. If schools wish to conduct instruction, but only after the HSA is administered, they may be able to start the instruction after completion of testing. Under this arrangement students would come to school at the beginning of the day if they are taking an HSA test. All other students would begin school around 11 a.m., when a shortened instructional day would begin for all students. Alternately, all students could complete a shortened instructional day with testing conducted at the end of the day. However, this may not be educationally sound as student fatigue becomes a concern. Additional buses would be needed and other transportation issues emerge in this scenario.

⁵ When 25% of students are completing the English 1 exam, 25% of classes will not be open. For example, a school may offer 12 English 1 classes throughout a 7-period day. If all students enrolled in these classes are completing the English test on Monday morning you will not have 12 open classes (because the same room may be used for English 1 in 5-7 periods). Instead you may free up about 3 classrooms in this example. These 3 classrooms could be used for testing, but where will the students enrolled in the remaining 9 sections of English 1 complete testing if instruction continues (of course these students completing English 1 will be absent from other classes scheduled during this testing time)?

Alternatives where no classroom instruction is conducted on days of HSA test administration

There is no school for students but all or most teachers will be involved in each HSA administration as test administrators or proctors. Only students who are completing the HSA test will report to school.

Will one or two HSA tests be administered per day?

Alternative Three - One test per day. This will result in 12 days for the HSA test administration and at least one make-up day during each test administration period. That translates to 13 days of testing for schools on a full-year schedule and 26 days for schools on a block schedule. If classroom instruction is canceled during test administration, this would eliminate 13-26 instructional days from a school's calendar. An illustrative schedule follows:

	<i>Week 1</i>	<i>Week 2</i>	<i>Week 3</i>
<i>Monday</i> -	English 1	Soc. Studies 1	Science 3
<i>Tuesday</i> -	English 2	Soc. Studies 2	Science 4
<i>Wednesday</i> -	English 3	Soc. Studies 3	Make-up Day
<i>Thursday</i> -	Math 1	Science 1	
<i>Friday</i> -	Math 2	Science 2	

Alternative Four - Two tests per day. This will result in only 6 days of test administration and one make-up day for each test administration period. That translates to 7 days of testing for schools on a full-year schedule and 14 days for schools on a block schedule. As noted above, if two tests are administered on the same day, MSDE will determine which combinations of exams will best minimize conflicts for students. In addition, these two exams could be offered at the same time, or one test may be administered in the morning and the second test administered in the afternoon. Two illustrative schedules follow (schools would actually have to determine the actual times administration begins):

<i>Week 1</i>	<i>AM</i>	<i>PM</i>
<i>Monday</i>	English 1	English 2
<i>Tuesday</i>	US History	World History
<i>Wednesday</i>	Algebra	Geometry
<i>Thursday</i>	Biology	Chemistry
<i>Friday</i>	Government	Physics
<i>Week 2</i>		
<i>Monday</i>	Earth/Space Science	English 3
<i>Tuesday</i>	Make-up 1	Make-up 2

<i>Week 1</i>	<i>AM (both exams begin at the same time and end at same time)</i>	
<i>Monday</i>	English 1	English 2
<i>Tuesday</i>	US History	World History
<i>Wednesday</i>	Algebra	Geometry
<i>Thursday</i>	Biology	Chemistry
<i>Friday</i>	Government	Physics
 <i>Week 2</i>		
<i>Monday</i>	Earth/Space Science	English 3
<i>Tuesday</i>	Make-up 1	Make-up 2

It appears that the latter two alternatives are more administratively feasible than the former alternatives, but the TASC and PSC groups can best determine this for MSDE. Instruction on testing days may be more easily accommodated in the latter than the former schedules.

5. Preparation Plus

The original conception of Preparation Plus has changed since it was initially presented as an option for MSDE. Since the winter briefings to MSBE, Science and Social Studies, which had originally desired to employ Preparation Plus, have changed to the Combination design. English is the only content area still proposing the Preparation Plus assessment design.

The new conception of Preparation Plus for English is to provide all students with a 60-minute preparation period prior to the test administration (either the hour preceding the test administration or the day prior). During this 60-minute period, the class teacher, designated test administrators, or certified substitute teachers will present stimulus materials. All teacher-student interaction will be highly scripted so there is no advantage to any class or student. The teacher may read extended instructions or present materials (e.g., videotapes, reading materials) during this time and any student work (e.g., reading, viewing, completing a work sheet) will be conducted during this 60-minute period. There was no general consensus about Preparation Plus among the TASC, and some members recommended that a state-wide feasibility study be conducted with actual school schedules prior to implementation. A few members of the PSC also recommended it might need to be reexamined; they noted that the preparation does not appear as relevant since few items are based on the preparatory activity.

The consensus of the MSDE design team is that the preparation piece is very important to both English content and test specification committees. They feel that Preparation plus can be implemented and that most issues and concerns still raised by CB/ETS have been addressed. They noted that this new model of Preparation Plus will allow students to

interact with and annotate text per HSA Core Learning Goals. The English Test Specifications Committee believes the new model will create a direct link between instruction and assessment. One hour of preparation time will precede the 3 hours of engaged testing. This new conceptualization of the Preparation Plus design reduces technical concerns about the English assessment but leaves some operational and feasibility issues to overcome.

For example, CB/ETS believe that among the unresolved issues surrounding the Preparation Plus model are: (a) operational feasibility for most school schedules, (b) degree of additional burden placed on schools to implement in a standardized manner, (c) amount of additional costs that will be incurred, and (d) technical and security concerns that may arise. It is quite possible that the Preparation Plus model may not meet all these challenges in a high-stakes testing program, and a feasibility study would be helpful to identify any remaining obstacles and additional costs before implementation is approved. However, it is also quite possible that Maryland educators will be able to address these challenges with adequate time, support and advanced planning. Following a feasibility study, a pilot implementation in a few schools with external auditors and evaluators would be advisable. Both the feasibility study and pilot implementation should be conducted prior to a state-wide no fault administration if possible.

5.1 *Rationale for Preparation Plus*

The English Core Learning Goals states that “Reading, writing, speaking, and listening require the learner to engage in preparatory activities and then to construct meaning, compose, and evaluate.” For this reason, the English Test Specifications Committee believes that these “preparatory activities” must be incorporated into the Preparation Plus phase of the HSA English tests:

“In the English classroom students interpret, generate, and evaluate texts. In Preparation Plus we envision that students will be able to interact with texts (print or non-print) such that they read or view them, interpret them, and evaluate them. We acknowledge that students possess a wide range of experiences with print and non-print materials and that ‘their experiences and backgrounds influence their understanding of the text’. Certain experiential variables, therefore, cannot be controlled in the Preparation Plus phase. To mitigate these variables as much as possible, the experience with text during the preparation phase should be contained and specifically scripted to provide students an opportunity to reflect not only on the text but also on the skills and processes they have mastered during their courses.

Reading a text in and of itself is not sufficient for understanding the richness of a text. Students need time to construct, examine, and extend meaning, including time for close reading, annotations, and reflective journals. Preparation Plus builds time and offers guided exercises so that students may engage in these enriching activities.

By providing preparation immediately prior to testing, this prep plus option forestalls out of class investigation. Prep Plus benefits all students, but most particularly LEP and Special Ed students by allowing more time for reflection and by shortening the daily block of time spent in test-taking.

Preparation Plus models the process of English instruction not only by providing opportunities for students to use before-, during- and after-reading strategies, but also by differentiating between surface reading and interpretation or analysis of a text.” (introduction to English CLGs)

5.2 Operational and Administrative Issues

The English Test Specifications Committee recommends that all English assessments be conducted over two days, with the preparation activity conducted the first day of testing. Section 4.3.1 of this report already addresses some of the operational, administrative, and technical burdens created if testing extends over two days.

However, even if English tests were administered entirely on one-day, the additional hour required for the preparatory activity may restrict options for scheduling two tests on the same day (again these are presented in section 4.3.1) and extend the test administrative schedule, reduce instructional days, and have some additional cost implications (development, printing of materials, use of teachers or substitutes for preparatory activity, etc.).

If preparation does occur on the day of the test, the preparation for all students could not possibly be conducted by their English teacher. Assuming English teachers may have more than one section of the same course, preparation would need to be conducted by other teachers for some classes. While the development of a rigid script should reduce the actual problems which may result from non-English teachers providing such preparation, there may be some perception of unfairness (students who are prepared by their English teacher vs. those prepared by the Health teacher) which will be more visible given the high-stakes use of the HSA. It is difficult to evaluate the feasibility of the Preparation Plus model until decisions are made on the overall scheduling of the HSA program and until feasibility studies and piloting of the model has been undertaken with actual master school schedules. This work should be conducted by MSDE as soon as possible.

CB/ETS does agree with the recommendations from the English Committee that if a preparatory activity is to be included it should: (a) be highly scripted, and (b) be conducted on the day of testing.

6. Equipment

This section deals only with equipment required to administer and complete the HSA and does not discuss the equipment needed to support the instructional components of the Core Learning Goals. This section is an overview of these requirements and a discussion of equipment common across more than one content area. Based on recommendation of the content area test specification teams, calculators can be required for 4 of the 12 assessments. The recommendations are:

Both math assessments	required.....	Graphing calculator
Chemistry	at least.....	4-function calculator
Physics	at least.....	Scientific calculator

There are a number of concerns and issues that MSDE must resolve concerning the use of calculators.

6.1 Can districts purchase the required types and numbers of calculators, train all math and science teachers in the use of calculators, and transmit a message that calculator use is required of all students for success in math and science by the beginning of the 1998-99 school year?

The most important requirement is to ensure that calculators are in the schools and teachers are prepared to provide instruction to students *prior* to the 1998-99 school year when the first state-wide field trials will be conducted. If Maryland educators have not received adequate training with the calculators that will be used for instruction, the validity and utility of data resulting from these field trials will be biased, and the tests may have to be delayed an additional school year. This staff training and equipment purchase would need to occur during the coming school year if the current schedule is to remain in place

Additional information on this issue is presented in Appendix G.

7. Test Development, Pre-Testing, and Scoring

The PSC had some recommendations concerning test development. When these comments addressed particular topics in a specific test, they were communicated to the Test Specifications committee, which considered how best to modify the test specifications to incorporate the suggestions, and these changes are reflected in the test specifications discussed in chapter III. The PSC comments that reflect general test concerns are addressed below.

7.1 Item Types

The PSC recommended that all multiple choice items have four answer choices with one correct response. Some members of the PSC also recommended a higher proportion of constructed response items relative to multiple choice items.

7.2 Test Content Reflecting Minority Topics

As mentioned in chapter 3, Test Specifications committees used as a working assumption that approximately 15% of test materials (textual, visual, and auditory) will portray multicultural and minority topics. The PSC recommended that test content include a minimum of 33% material involving people and topics (passages, textual materials, names) portraying minority interests or perspectives. While it may be possible to attain the objective recommended by PSC, it would be more efficient to strive for this level, while retaining the lower minimum (i.e., 15%) for each test. Requiring 33% of all items to portray minority topics is likely to constrain other areas of item development and test assembly that will increase costs and require significantly more effort and time to generate acceptable items. The sensitivity review process should address any concerns about the nature of item and test reviews (see Appendix F for sample procedures).

7.3 Pre-Testing

Some members of the PSC felt that pre-testing of multiple choice items should not be conducted as part of the actual HSA tests, while other members felt that such a pre-testing design was permissible if the number of items was relatively low. The proposed pre-testing methods described in Appendix B does assume that a small number of multiple choice items that do not count toward a student's actual score will be contained in all tests to permit equating of test forms and to allow the development of new forms of each test.

7.4 Review of test items and content

This issue concerns under what conditions parents, policymakers, or other citizens may review test items, correct responses, and completed forms of students participating in the HSA. The PSC recommended that security of all materials be maintained at all times. Any reviews should be conducted under secure conditions (e.g., supervised reviews, time limit for review, and prohibition of copying). If parents or citizens are to have access to test materials, there should be defined procedures in place, and copying of information should be prohibited unless the item or test form is disclosed. In addition, a sensitivity review team should be established to review potential items, and the PSC recommended members of this group come from MSDE staff as well as non-MSDE staff (e.g., special education teachers, linguistic reviewers, classroom teachers, testing experts, university faculty, members of the business community, parents and staff). Finally, PSC recommended all members of the sensitivity review team have some knowledge in the relevant content area and with the Core Learning Goals.

7.5 *Disclosure Of Forms*

The PSC did not recommend that a test form be disclosed to the public until there is a surplus of items. CB/ETS agree that releasing a disclosed test form is appropriate when the pools of test questions are sufficiently rich to enable this release.

However, samples of items should be disclosed as soon as data can be gathered to indicate that they are representative of the kinds of items to be included in the HSA tests.

8. Technology

The PSC recommended that a higher level of technology be employed with the HSA. They believe that word processors should be employed for essays and that Internet and computerized testing be used for all tests, etc. They noted that if these technologies could not be employed initially, they should be phased in within the next few years.

These are important areas for MSDE to consider for the future, but there would be many concerns with introducing any of these into the HSA in the next few years. The costs for computer-based testing, word processing, and electronic dictionaries may be prohibitive for another few years, given the costs for development and operation of the HSA as currently proposed.

9. Staff Development

Three distinct types of staff development will be needed for Maryland educators over the next few years to ensure the successful implementation of the HSA. First, Maryland educators will require staff development to assist them in understanding, operationalizing and implementing the Core Learning Goals. Teachers will need to understand the goals, expectations and indicators relevant to their content area. In addition, teachers will need staff development and instructional materials to assist in implementing the Core Learning Goals through innovative pedagogy and other instructional practices. Educators in a few districts report some limited staff development in these areas, but have generally expressed concerns that the majority of teachers do not presently have the prerequisite knowledge, skills, instructional materials and support to implement the Core Learning Goals. Staff development in this area appears to be unsystematic. Staff development must also migrate to middle school and upper elementary level educators (not reside only at the secondary level) if Maryland's students are to become proficient in these high standards by high school. If the lack of systematic staff development remains unchanged by 1998-99 (the year of state-wide field testing), the implementation of the HSA may be flawed because students will not have had an opportunity to learn the knowledge and skills measured in the HSA. For example, a description of the need for teacher training in the use and instruction of students in using a graphing calculator was illustrated earlier. If teachers lack the requisite knowledge in this domain, but students are held to this standard (through assessment items), the test will create an unfair burden on students.

Professional standards of practice suggest that MSDE should conduct a survey (or use other means) to ascertain that schools and educators are implementing the Core Learning Goals (and students have had multiple opportunities to learn these goals) prior to the use of the HSA for any individual level or school based decisions.

Second, Maryland educators will require staff development on the specific operational issues involved in a large-scale assessment system such as the HSA. Specifically, educators will need staff development that addresses:

- the content and format (tasks, items) contained on the HSA tests in their content domain, including illustrative items and rubrics
- issues of test security, instructional materials that can and cannot be displayed, acceptable test familiarization practices, methods of reducing anxiety and increasing student comfort levels, etc.
- issues of comparable administrative conditions for standardized assessments
- an overview of the design and uses of the HSA, discussion of the implications for students, teachers, schools and districts, and the HSA schedule
- issues of remediation and retesting for students who fail the HSA
- other issues as they emerge throughout test development
- interpretation of scores
- use of results for curriculum, instruction and remediation

Such training will need to be ongoing over the course of the HSA and should begin during the 1997-98 school year. This training is essential to ensure teachers are familiar with the format, uses, and results from HSA and can explain and interpret results for parents and students.

Third, Maryland educators who will serve as test administrators and proctors will need detailed training on the specific tests when they are developed for the state-wide field test in 1998-1999, and each subsequent year. MSDE should develop a standard set of training materials that will address the specific issues involved in administration, maintenance of security, distribution, and collection of the HSA.

Additionally, if teachers are to be used to score the HSA assessments, additional training will be required before teachers can be qualified as readers. Ongoing evaluation and retraining of readers will be required, and those readers unable to meet and sustain appropriate levels of scoring accuracy will need to be dismissed from the readings. The TASC recommended that all staff development be in place prior to the first state-wide field testing of the HSA.

10. Scoring And Score Reporting

Scoring and score reporting involve many complicated issues. Issues relative to scoring brief and extended constructed response items are discussed in Appendix B. Still other items are presented below.

10.1 *Scoring And Appeals*

The PSC recommended that MSDE establish a well defined appeal process for students and parents who wish to question a score. The process should be the same for all tests and should be communicated widely. The process should consist of: (a) an automatic review of any score that is near the decision point, and (b) an appeal process for student and parent requests.

10.2 *Proficiency Levels*

The PSC recommended that proficiency levels (not pass/fail) be used in scoring students' assessments, but took no position on how scores should be reported. This recommendation is consistent with the decision made by the MSBE. Proficiency levels can provide more useful information to test users and students. However, with each additional proficiency level, there is a need to develop clear and precise cut-points for students above or below that level. As the number of levels increases, the number of test items close to that level must also be increased. Since MSDE has identified three possible uses for scores, it would be preferable to have only three or at most four proficiency levels because of the demands on test construction and test length. Five or more proficiency levels would likely require more multiple choice items and longer tests if high reliability of scores is to be maintained (see Appendix B).

10.3 *Schedule for Score Reporting and Scoring*

The interim report provided an extended discussion of two potential schedules for test administration and their implications on score reporting (see pages 18-21 and pages 46-48 of that report). In the interim report, we noted that participants in all public engagement activities had emphasized that:

- assessments must be administered as close to the end of the course as possible, and
- results must be available to students and schools before the end of the term.

We also noted that these two conditions appear to be mutually exclusive and have not been successfully implemented in any large-volume, paper-based, high-stakes assessment program employing constructed response tasks (such as the HSA). We estimated about 7-9 weeks between the test administration and score reporting for the combination assessment design, under the best conditions. Meeting this estimate for score turnaround will require substantially increased financial and human resources from the state and the schools, as well as the successful resolution of a number of very critical issues. Thus, there is a significant trade-off between minimizing score turnaround time and the possibility of including performance assessment components in each HSA instrument. A more obvious trade-off is

between end-of-course administration, as proposed, and having scores available for counseling, placement, and remediation of students who fail the test must be weighted.

A related issue is the timing of reporting scores to schools, students, and parents, e.g., must the scores be returned before the end of the semester or course? In planning the schedule for an assessment program, one can work backward by first specifying when the test scores must be reported and using that information to determine the latest date for administration. The high-stakes impact on individual students is cited as a major reason that scores must be returned by the end of the term. However, there are other reasons:

- To enable schools to plan remediation that may be required during the summer or in the following semester for students who fail, as well as to inform course placement decisions.
- To build a master school schedule.
- To provide scores to seniors before graduation.
- To provide scores to students and schools while they have meaning. Quick turnaround is viewed as essential for ensuring that the assessments are viewed as important levers of reform in the schools.

The time required for scoring and processing the assessments creates a strategic conflict for many constituency groups that both want scores before the end of the course and want the assessments administered as close to the end of the course as possible. It is clear that both of these demands cannot be met via a large-scale, high-stakes, paper-and-pencil testing program.

However, some educators have argued that administering the assessments well before the end of the course would have a beneficial impact on the curriculum by creating a period of time when teachers must go beyond the Core Learning Goals. The perceived benefit of an earlier administration stems from a concern that the curriculum could become synonymous with the Core Learning Goals (CLGs) and not include the additional, but not necessarily common, topics and skills that were envisioned by the content teams that authored the CLGs.

Given the practical limits on turnaround time, there are two distinct alternatives⁶:

- administer the HSA part way through the course and report the results just prior to the end of the course (relaxing the association between assessments and conclusion of the course); or
- administer the HSA at the end of the course and report the results several weeks after the end of the semester.

⁶ These are the only alternatives, if MSBE retains the requirement to have ten separate end-of-course tests. Earlier discussions about moving to an end-of-program assessment system or introducing four content tests (which might be administered mid-way through the next course level) would provide additional alternatives that could address some of the concerns about score reporting.

The first option provides the possibility of teachers using HSA results as part of a course grade as well as facilitating the placement of students into sequence courses or remediation. It may, as has been argued, encourage teachers and schools to implement a more comprehensive curriculum that extends the subject matter beyond the CLGs.

The second option provides the students with maximum time to master the CLGs covered in a particular assessment. This option would more easily accommodate the use of Maryland teachers to score the HSA.

These two alternatives are illustrated in Table I.

TABLE I - Illustrative schedules for administering and reporting HSA

<i>HSA Administered</i>	<i>HSA Results Reported</i>
Option 1 - Scores Reported Before End of Course/Semester	
<i>1A- Full-Year Program or Spring Semester Block</i>	
Administration: 4/5 - 4/15	Scores Reported: June 7
<i>1B-Fall Semester Block</i>	
Administration: 11/1 - 11/10	Scores Reported: Jan 10
Option 2 - Scores Reported After End of Course/Semester	
<i>1A- Full-Year Program or Spring Semester Block</i>	
Administration: 5/24 - 6/2	Scores Reported: July 28
<i>1B-Fall Semester Block</i>	
Administration: 1/5 - 1/15	Scores Reported: March 11

Under either scenario there are problems and compromises which must be considered.

<i>Scores Reported Before the End of the Course</i>	<i>Scores Reported After the End of the Course</i>
<ul style="list-style-type: none"> • Administer tests early in course (Nov. 1, Apr. 5) • Schools on block schedule most adversely affected by schedule • Less coverage of CLGs • Scoring of HSA by Maryland teachers precluded • Have scores to aid in placement and remediation decisions • Have scores if school wishes to use them for course grades • Scoring schedule is tight and there is a potential that scores may not be ready in time at outset 	<ul style="list-style-type: none"> • Administer Tests Later in Course (May 24, January 5) • Schools on block schedule somewhat less affected • More coverage of CLGs • Scoring of HSA by Maryland teachers possible for spring form only • Scores may be late to aid in placement and remediation • No information for course grades • Scoring schedule can be a bit more flexible and scores will be ready by mid-summer

No recommendations were made by either the TASC or the PSC regarding whether tests should be administered earlier in the semester to provide scores prior to the end of the course, or whether they should be administered later in the course precluding score reporting prior to the end of the year. Either scheduling model can be made to work with time and coordination.

Since the interim report was issued, there appear to be some possible technological enhancements on the horizon that promise to reduce the time between administration and score reporting for constructed response tasks. Image processing of textual and graphic materials can be done by scanning students responses at a central district location. Several test developers are scoring student responses that have been scanned in this way through either centralized scoring locations or remote scoring processes. Several organizations are using this technology currently, although it still may not have adequate psychometric precision for the type of high-stakes individual decisions proposed for the HSA. However, the estimated 7-9 week interim between administration and scoring could be reduced by approximately 30-50% if the psychometric and operational issues can be resolved in the next two years and such assessments become more affordable. Based on current technologies, we estimate that image processing of student responses and central or remote scoring networks may be developed to the stage where they can be applied to the HSA in the next three years. MSDE should monitor advances in this area. Current and future image processing and scoring capabilities of bidders for the HSA scoring proposal should be considered in this award.

10.4 Teacher vs. External Scoring

Scoring of constructed response items will require several days or possibly weeks of scoring. First, readers must be trained in scoring the papers using generic rubrics. Training may last between 5 -15 days depending on the volume of readers used. Because of the high stakes associated with the HSA and the accompanying need for quality control in scoring, we do not recommend that all Maryland teachers score papers. Teachers should not score papers from students in their own classes because of the need for both high quality scoring and widespread credibility of the scoring. However, even if teachers were to read papers from another school, it is not likely that all Maryland teachers will have the appropriate level of skill and experience required for scoring these assessments. It is more efficient and technically more sound to have a smaller number of readers score a moderate to large proportion of papers than to have a large number of readers score a small proportion of papers. In this way, standards for scoring can be retained and evaluated throughout the scoring to minimize rating biases and rater drift.

If a subset of Maryland educators are to score the HSA papers they will require a minimum of 5 full-days (probably 10-15 days) each to complete the process and set the standards for proficiency levels. No district may be able to excuse large

numbers of teachers twice each year following each administration. Thus, external scoring of the HSA may be the most efficient alternative. Again, no specific recommendations regarding this issue were made by either the TASC or the PSC.

10.5 Local, Regional, and Central Scoring Options

This issue concerns who will score the HSA tests, where they will be scored, and what options exist for scoring by Maryland educators (also see discussion in chapter II of this report) . There are three basic options for where the HSA tests may be scored:

- Central Scoring—all papers are returned to one location where they are scored by readers assembled at that location. For example, external readers or Maryland teachers would convene at one location and score all English 1 tests. Additional central scoring for other tests (e.g., math, science, social science) may be conducted at other sites, but this would still be defined as central scoring.
- Regional Scoring—papers from one test are shipped to several regional locations for scoring by external readers or Maryland educators. This requires similar training, standards, sample papers, and adjustments across sites to ensure comparable standards.
- Local Scoring—each district scores tests from its own students. A separate audit of 5-10% of papers could subsequently be conducted by MSDE to ensure rating standards, but when discrepancies in ratings are discovered, there are no mechanisms to adjust individual papers.

Central scoring is used when a high degree of reliability and consistency across raters is required, as is the case with high-stakes assessments producing individual scores. Regional scoring is an option that can be considered, but maintaining consistent scoring across multiple sites will be technically and logistically challenging (papers will be shipped to the wrong sites, greater risk of error and missing papers) and probably offer little above central scoring. Local scoring is simply not realistic as the primary vehicle for arriving at individual scores that will be used for high-stakes decisions. Local scoring could be an option for schools that wish to provide staff development for their teachers and possibly have scores used for local course grades. Then, MSDE would need to obtain the papers for rescoring at a central scoring. For only high-stakes, school-level uses or for low-stakes (e.g., indicators of relative strengths and weaknesses) individual uses, local scoring can be an option with a central audit of scoring quality by the state.

If consistent scoring standards across the state are to be achieved, central scoring is required. If MSDE is willing to accept inconsistencies in scoring across districts or is willing to permit each district to establish its own standards based on generic rubrics and proficiency statements, then local scoring, with an audit process, may be

attempted. However, individual high-stakes uses of resulting scores will not likely be psychometrically defensible under this scenario and may not be legally defensible if students are penalized for performance. Again, MSDE must choose among the benefits of having local scoring conducted by Maryland educators and the extreme high-stakes uses proposed for these tests. If standardized high-stakes uses are proposed for tests, than standardized administrative conditions and scoring are required.

There are generally two options for who may score these assessments:

- a subset of Maryland teachers in a subject area may serve as scorers
- external scorers may be hired to conduct the scoring through a vendor or a state-established scoring center (in collaboration with a university)

It is much more feasible to use Maryland teachers as scorers if the HSA will be scored at the end of the school year (as is the case with the MSPAP). Teachers could then conduct the scoring under MSDE supervision or be hired by a vendor (again, using the MSPAP model). If scores are desired before the completion of the school year, then having Maryland educators conduct the scoring is more problematic. First, there is the concern about the high volume of tests anticipated across the state. Next, each test will have multiple CR items requiring at least two scorers and consuming substantial time to score reliably. Third, even with low volumes, it will require at least 10-15 days for a substantial number (10-20%) of all Maryland teachers in a subject area (e.g., English) to be trained in scoring papers, conduct the scoring, and have some involvement in pre-scoring and post-scoring activities (e.g., selecting benchmark papers, establishing proficiencies). If scoring occurs during the school year, how can so large a proportion of teachers be excused from instructional duties for two or more weeks? Under the best scenario, the winter administration (for schools on semester block schedules) will present a formidable challenge for using Maryland teachers.

Given the high volume and multiple administration dates, Maryland's most feasible options may involve using external vendors to conduct the actual scoring, with a small percentage of teachers involved in selecting benchmarking papers, overseeing the scoring, and establishing proficiencies.

Because all tests include a section of MC items, booklets must have ID numbers on each page to permit them to be physically separated for scoring. The MC portion of booklets would be scanned by MSDE or a vendor, while the CR items are prepared for human scoring. Earlier in this report there is a brief discussion of the use of image processing to reduce the cost and time required for scoring the CR items in the future. The interim report also discusses the potential of computer-based administration and scoring well in the future. However, the latter option is not a viable alternative in the next five years, as explained in the CB/ETS interim report.

11. Retesting And Remediation

The PSC made several recommendations regarding these issues. The PSC recommended that appropriate assistance be made available by the state and districts for all students failing the HSA. Should students decline this assistance, parents should be required to sign a parental agreement form. The PSC also recommended that districts begin designing appropriate assistance programs for students who fail each HSA test immediately. MSDE should provide general guidelines and identify promising practices to aid districts in developing these programs. Appropriate assistance may often be in the form of review or remediation instruction that districts may provide during the summer, but alternatives such as special courses, tutoring, after school or weekend instruction (provided at the school or county level) should be considered. Finally, the PSC recommended a third administration during the summer be provided to accommodate students who complete the remediation and wish to retake the test immediately.

Generally, the issues raised here—appropriate assistance for students failing the HSA—have simply not been adequately addressed and must be considered well before there is any use of the HSA for high-stakes individual uses. If adequate opportunities and funding to support these opportunities are not in place for students who fail the HSA, high-stakes uses will not be professionally, ethically, or legally defensible.

12. Costs

This issue concerns who will pay for the various types of assistance and operational components associated with the HSA. The College Board and ETS have no recommendations on these issues, but will report the recommendations from the TASC below⁷:

- State and local support are required for the purchase of all equipment, staff development, and support for the administration of assessments. This support is required prior to the commencement of school in the year of the first state-wide no-fault administration. If local districts are to purchase equipment, they will need funding and detailed information one year in advance of this (or by January 1998).
- A minimum of one dedicated staff person will be required at each school, and it is expected that local districts will be responsible for these costs.
- A number of special studies must be designed (e.g., speededness, varying number of MC items which can result in a reliable test, calculator differences) and incorporated into the upcoming no-fault administrations and any pilot testing that is conducted. All studies must include representative groups of

⁷ Summarized from the notes of the May 12, 1997 meeting of the Test Administration Specifications Committee.

students from special needs and LEP populations, and include geographically diverse schools.

- Local districts may be required to pay for additional substitute teachers if they would be needed for implementation of Prep Plus and the assessments. Detailed information on the administrative requirements is needed to help locals prepare for these requirements.
- Local districts will also be responsible for developing remedial classes and other appropriate programs for student who do not pass the required tests.
- The state will be required to provide financial support for the development of all tests, printed materials describing the tests, scoring and score reporting, special research studies, and a prototype test that can be shared with educators in the next year. State will assist locals in supporting remediation, equipment purchases, and staff development expenses. Local districts will support staff and substitute teachers required for administration and instruction, and additional staff required to manage the school's assessment scheduling and administrations.

13. Transition

This is an important issue which discusses how MSDE might transition from the current use of minimal competency Functional Tests to a fully operational HSA system that would be used for high-stakes purposes such as high school graduation. There are a variety of ways for introducing a high-standards and high-stakes assessment program as Maryland desires to do in the next few years. However, some of these methods may be more politically acceptable, more feasible, more cost effective, more technically sound and more successful, in the long run, than other methods. The interim report provided an extended discussion of state efforts that have failed for a variety of reasons. We attempt to provide some recommendations to facilitate this transition. Many of these recommendations may not be popular with specific constituency groups. Many of these recommendations may appear to be creating unnecessary delays or impediments with the overall goal to move to high stakes, high standards and high consequences for all students in Maryland as soon as possible. However, we believe that issues of implementation and transitioning should be considered seriously by MSBE to ensure the ultimate technical, operation, political, and economic success of the HSA in the long run.

CB/ETS recommends that MSDE review the information presented in the interim report—that information, much of which supports these brief recommendations concerning the transition and introduction of the HSA, is not repeated again in this report. Based on lessons learned from other testing programs, we provide a few recommendations concerning the initial introduction or transition to the HSA tests. Again, these recommendations concern initial implementation (not final operational specifications) of

the HSA and attempt to minimize technical and administrative problems that could threaten a new assessment system, as well as increase the acceptability of the assessments to the public and Maryland educators.

13.1 *Begin with fewer tests if necessary*

Begin with only a few tests and introduce additional tests iteratively. This will provide additional time for the necessary funds to be allocated to the tests, provide a longer learning curve for students, parents, educators and the public to become aware of the tests, and reduce the level of effort for schools and MSDE (and related burdens to students). Going too far, too fast has serious potential negative consequences for students, the quality of education, and the future success of the entire HSA.

13.2 *Determine need for, expected volume, and additional costs for physics and chemistry tests*

Reconsider development of chemistry and physics end-of-course tests. Four separate science tests will be available, but students will only be required to complete two of these. CB/ETS believe that most students may opt for the biology and environmental/earth science tests because: (a) they will often complete these courses prior to chemistry and physics in most districts, and (b) exams in these subjects may be perceived as “easier” than those in physics and chemistry. These two factors, plus the smaller percentage of students enrolling in chemistry and physics may mean that developing tests in these subjects will be less cost effective than other areas. If less than 10% of students complete both tests (electing to complete the other science tests), it may also be difficult to equate forms. MSDE should consider conducting a survey of district enrollment patterns and course sequences to determine demand for physics and chemistry tests before expending significant resources in test development. The science educators involved in the test design would strongly disagree with the CB/ETS recommendation since their objective is to reform and enhance science curriculum and standards in Maryland.

13.3 *High standards first, high stakes later*

It may be unfeasible to introduce high stakes with high consequences at the same time. If this situation exists, MSDE should consider first establishing high standards for the test and then gradually increasing the stakes associated with the test scores. This, we believe, is a better alternative than introducing high stakes initially, and gradually increasing the standards (the approach New York state is initially pursuing). Translating standards (what students must be capable of doing) to the public will be difficult. It often takes many years before students, parents, educators, business leaders, and other citizens realize and can themselves articulate the high standards and expectations for students. If these standards change each year, this process will be much more difficult. Raising the bar iteratively may

ultimately be confusing if MSBE desires acceptance and understanding of the high standards. As with MSPAP, a substantial proportion of students may not reach these high standards initially despite any number of opportunities to retest or intensive remediation interventions. This would be problematic if high school graduation were contingent on the high standard. Rather than punish students who do not initially meet these high standards, Maryland can reward the students who do meet the standard. There are a variety of rewards that could be considered that range from issuing an endorsement on students' diplomas or providing financial assistance for students entering Maryland's public higher educational system to working with businesses to use endorsements for hiring and entrance into training programs.

13.4 *Local choice and flexibility vs. standardization*

While Maryland educators highly value flexibility and local choice on all operational issues related to HSA design (e.g., modules), scoring (local teachers), and administration (e.g., local procedures, different times for administration and breaks), high-stakes testing requires standardization if results across students, schools, and districts are to be comparable. Even if test items are carefully reviewed, unfairness can result when there are variations in the items (e.g., content, difficulty) or administrative conditions (e.g., amount of time permitted to complete the tests, quality or type of tools permitted, preparation activities, security procedures invoked), or scoring (e.g., variability among scores in the standards used for scoring). Strict adherence to content specifications will make different forms of the same test comparable, so students should not care which form of the test they complete. When variations are permitted in test items, administrative conditions, or scoring, this is no longer the case. Comparability cannot be ensured, and variations will introduce unfairness in the resulting scores. If local autonomy of testing is the ultimate value, then greater flexibility can be accommodated with a diagnostic assessment system. If accountability and student and district comparisons are the central goal, then standardization and accurate scores will prohibit some of the preferences associated with local control of testing. If both objectives are equally valued, then it is likely the HSA will fail to meet one, or possibly both objectives (also see chapter II).

13.5 *Introduce fewer decision points and proficiency levels initially*

Introduce as few decision points and proficiency levels as possible to ensure that the tests can precisely measure those decision points associated with high stakes. Multiple proficiency levels provide valuable diagnostic information, but will not be supported with adequate numbers of test items if four or more cut points or proficiencies are desired. The test specifications do not include sufficient numbers of items to permit use for multiple decision points and provide highly reliable diagnostic information.

13.6 *Do less better and win acceptance and approval from the majority of key stakeholders rather than risk serious problems with such an ambitious assessment system*

Initially, “do less better” before attempting to do more with the HSA. Be sure the HSA is administratively feasible, professionally acceptable (technical specifications) for its intended purposes, legally defensible, and economically affordable by limiting the number of “variations,” “uses,” and “bells and whistles” associated with the tests. Introduce additional uses only after the HSAs are proven to be useful and valid for the primary uses. Consider introducing greater flexibility for administration and scoring only after it has been demonstrated that the HSA can be successfully implemented (administered and scored) in schools with a minimum of options and choices. Consider having transfer students, home-bound students, and private school students complete the tests only after all Maryland public school students can successfully be accommodated. Examine options such as computer administration of tests, variable make-up options for Preparation Plus, and local scoring only after it has been demonstrated that the HSA can be successfully administered and scored using current technology and resources. It is easy to promise more than can be delivered with any assessment. By promising less, initially, and doing more, assessments are less likely to disappoint and disillusion key constituency groups that are required to sustain their support.

REFERENCES

- Allen, N. A., Kline, D. L., & Zelenak, C. A. (1996). *The NAEP 1994 technical report*. National Center for Education Statistics.
- Camara, W.J., Kimmel, E.K., and colleagues (1997). *High School Assessment Design: Interim Report to the Maryland State Board of Education*. New York: The College Board.
- Carlson, J. E. (1968). *Effects of differential weighting on the inter-reader reliability of essay grades*. Unpublished Dissertation, University of Alberta.
- College Entrance Examination Board (1988). *The college board technical manual for the advanced placement program*. New York: College Entrance Examination Board.
- Gulliksen, H. (1987). *Theory of mental tests*. (Reprint of work originally published in 1950). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kolen, M. J. & Brennan, R. L. (1995). *Test equating methods and practices*. New York: Springer.
- Wang, M. W., & Stanley, J. C. (1970). Differential weighting: a review of methods and empirical studies. *Review of Educational Research*, 40, 663-705.

Appendix A

Additional Test Specification Recommendations

ETS considers some of the recommendations from the Test Specifications committees to be premature in the absence of information and data to inform the decisions. It is the opinion of ETS that recommendations which specify detailed test assembly, item scoring, weighting of item types, or test administration instructions should not be acted upon until further development and no-fault administration has occurred. Test assembly guidelines should be based, in part, on the difficulty level of individual items and item types.

Weighting of item types should be based on psychometric characteristics of the item types. Scoring issues should be addressed by the scoring advisory team which will provide the scoring tools based initially on rubrics outlining the correct answer and further refined after no-fault and/or pilot data on the questions has been gathered. Test administration instructions are within the purview of the Test Administration Specifications committee.

The sections below contain recommendations from various Test Specifications committees which should be considered fully at a later time.

TEST ASSEMBLY

MATH 1

Recommended Test Configuration:

It is recommended that the test be administered over 2 days, as follows:

Day 1:

Directions
3 ECRs
5 BCRs
30 SRs

Total time: 96 minutes

Day 2:

1 ECR
6 BCRs
10 SPRs
30 SRs

Total time: 84 minutes

The following principles should be considered when configuring the test:

1. More constructed response questions should appear earlier in a part so that students are fresh when they attempt them.
2. The SPRs should be kept together within a part.

[ETS comment: The advice about test assembly is based on two days of testing. ETS recommends that the test be given on one day. We also recommend that no-fault data be used to inform decisions about placement of item types.]

MATH 2

Recommended Test Configuration:

It is recommended that the test be administered over 2 days as follows:

Day 1:

Directions
2 ECRs
6 BCRs
25 SRs

Total time: 95 minutes

Day 2:

1 ECR
3 BCRs
10 SPRs
35 SRs

Total time: 85 minutes

The following principles should be considered when configuring the test:

1. The ECRs and BCRs for the same expectation should be in the same part of the test.
2. More constructed response questions should appear earlier in a part so that students are fresh when they attempt them.
3. The SPRs should be kept together within a part.

[ETS comment: The advice about test assembly is based on two days of testing. ETS recommends that the test be given on one day.]

WORLD HISTORY

Recommended Test Assembly and Administration:

Section I:	SRs	45 minutes
	BCRs (4)	20 minutes
Break		10 minutes
Section II:	ECR	35 minutes
Break		10 minutes
Section III:	SRs	45 minutes
	BCRs (4)	20 minutes
Testing Time		165 minutes
Time for Instructions:		15 minutes
Break Time:		20 minutes
Total Time of Admin:		200 minutes

[ETS comment: The advice about test assembly is premature. We recommend that no-fault data be used to inform decisions about placement of item types.]

U.S. HISTORY

Recommended Test Assembly and Administration:

Section I:	SRs	45 minutes
	BCRs (4)	20 minutes
Break		10 minutes
Section II:	ECR	35 minutes
Break		10 minutes
Section III:	SRs	45 minutes
	BCRs (4)	20 minutes
Testing Time:		165 minutes
Time for Instructions:		15 minutes
Break Time:		20 minutes
Total Time of Admin:		100 minutes

[ETS comment: The advice about test assembly is premature. We recommend that no-fault data be used to inform decisions about placement of item types.]

GOVERNMENT

Recommended Test Assembly and Administration:

Section I:	SRs	45 minutes
	BCRs (4)	20 minutes
Break		10 minutes
Section II:	ECR	35 minutes
Break		10 minutes
Section III:	SRs	45 minutes
	BCRs (4)	20 minutes
Testing Time:		165 minutes
Time for Instructions		15 minutes
Break Time		20 minutes
Total Time of Admin:		200 minutes

[ETS comment: The advice about test assembly is premature. We recommend that no-fault data be used to inform decisions about placement of item types.]

ENGLISH

Recommended Test Assembly:

For English 1 and English 2:

Day One:

60 minutes preparation period

45 minutes BCRs

15 minutes SRs

Day Two:

50 minutes SRs

30 minutes ECR

25 minutes Srs

For English 3:

Day One:

60 minutes preparation period

45 minutes BCRs

15 minutes Srs

Day Two:

40 minutes SRs

30 minutes ECR

35 minutes SRs

This template is recommended unless data suggests otherwise.

[ETS comment: This assembly is based on two days of testing. ETS recommends that the tests be given on one day. We also recommend that no-fault data be used to inform decisions about placement of item types.]

ITEM SCORING

ETS comment about advice which follows: Advice about weighting and scoring is premature. We recommend that specific scoring rubrics be developed after items have been written. We also recommend that no-fault data be analyzed to determine the most appropriate weighting for the various item types.

BIOLOGY

The Biology Test Specifications committee has recommended activity-specific scoring for the constructed response questions.

CHEMISTRY

Suggested scoring for CR items:

The rubrics should indicate that evaluation is based on both the process skills and the concept knowledge exhibited in the response.

BCRs

- 0 = no understanding
- 1 = little understanding
- 2 = basic understanding
- 3 = full understanding

ECRs

- 0 = no understanding
- 1 = little understanding
- 2 = basic understanding
- 3 = full understanding
- 4 = full understanding plus
extension of ideas

WEIGHTING

WORLD HISTORY

The Content Specifications Committee recommends the following weights:

Across the entire test:	SR	50%
	BCR, ECR	50%
Among CR items:	BCR	32%
	ECR	18%

[ETS comment: The advice about weighting is premature. We recommend that no-fault data be analyzed to determine the most appropriate weighting for the various item types.]

CHEMISTRY

The specifications committee recommends that the sections be weighted so that selected response counts for 50% of the final score, and constructed response counts for 50% of the final score. The contribution of the two types of CR to the CR score should be proportioned to time, so that the ECRs contribute 40% and the BCRs 60%. The percents noted earlier for each question type are based on these recommendations.

[ETS comment: The advice about weighting is premature. We recommend that no-fault data be analyzed to determine the most appropriate weighting for the various item types.]

PHYSICS

The specifications committee recommends that the sections be weighted so that selected response counts for 60% of the final score, and constructed response counts for 40%. The contribution of the two types of CR to the CR score should

be proportional to time, so that the ECRs contribute 60% and the BCRs 40%. The percents noted for each question type in the first table are based on these recommendations.

[ETS comment: The advice about weighting is premature. We recommend that no-fault data be analyzed to determine the most appropriate weighting for the various item types.]

U.S. HISTORY

The Content Specifications committee recommends the following weights:

Across the entire test:	SR	50%
	BCR, ECR	50%

Among CR items:	BCR	32%
	ECR	18%

[ETS comment: The advice about weighting is premature. We recommend that no-fault data be analyzed to determine the most appropriate weighting for the various item types.]

GOVERNMENT

The Content Specifications committee recommends the following weights:

Across the entire test:	SR	50%
	BCR, ECR	50%

Among CR items:	BCR	32%
	ECR	18%

[ETS comment: The advice about weighting is premature. We recommend that no-fault data be analyzed to determine the most appropriate weighting for the various item types.]

TEST ADMINISTRATION

MATH

Recommended Test Assembly:

The tests should each be taken in two parts; each part should contain some multiple-choice and some constructed response. Questions should be ordered by difficulty from easy to hard. The Content Team recommends that the tests be constructed to motivate students to continue to work throughout the test.

[ETS comment: ETS recommends that testing should be done on one day.]

ENGLISH

The Preparation Plus model calls for the English test to have the following shape:

Day 1:

60 minutes preparation time

60 minutes testing time

5 minutes for directions

Day 2:

105 minutes testing time

10 minutes for directions

[ETS comment: ETS recommends that testing be done on one day.]

Appendix B

Technical Specifications

1. Introduction

The Maryland High School Assessment (HSA) is designed to yield information that will be used for making important decisions. The assessment instruments (tests) must, therefore, meet very rigorous technical specifications.

2. Validity

Validity of an assessment instrument refers to issues about whether the instrument truly measures the knowledge, skills, etc., which it was designed to measure. In this section are addressed some general validity issues for the HSA and two specific types of validity that are most relevant.

2.1 General Considerations

Validity must be considered in the context of the assessment program. For the HSA there are three elements of the context that require attention: basis in the core learning goals, consistency across school districts, and purposes of the program.

2.1.1 Basis in Core Learning Goals

Because the content specifications for each test are based on the Core Learning Goals, a good foundation exists upon which to build the content validity of the tests. With appropriate development and review procedures for the test items and materials, it should be possible to establish that the tests are measuring content from the Core Learning Goals, providing one essential link in establishing evidence of content validity of the HSA.

2.1.2 Cross-district Concerns

The second linkage required for content validity is established by demonstrating that curriculum and instructional practices across districts properly include the Core Learning Goals that are contained on the HSA. Such evidence of student opportunity-to-learn can be established through curriculum audits, teacher surveys, and a variety of other means. This evidence should be in place prior to the operational administration of the HSA.

2.1.3 Purposes of the HSA

Validity in testing also must be considered in relation to the uses to be made of the tests. Projected uses of the HSA include certification of students at various levels of mastery (at least two levels: pass/fail) and to help determine measures that may be taken to help students not initially passing to reach that level. In order for decisions based on HSA results to be valid, the tests must be constructed to represent the content specifications precisely.

2.2 Relevant Types of Validity

Measurement experts have identified a number of different types of validity that should be addressed in assessment programs. The types that are most relevant for the HSA are content validity and decision validity.

2.2.1 Content Validity

Each item on each test must be constructed to assess one or more elements of the content specifications. Additionally, the set of items on any given test must closely match the overall content specifications. Each test must be demonstrably related to skills and abilities that the test is designed to measure. Thus, each form of each test in the HSA battery must be carefully evaluated for coverage of each element of the content specifications according to specified percentages. This evaluation should be done by the test developers.

2.2.2 Decision Validity

The important decisions about passing HSA exams, and about remediation for students not passing, must be based on valid information derived from the test scores. If remedial decisions are to be based on classifying individual students according to defined levels, for example, the validity of classification into levels (See section 13, Multiple levels of performance) is of concern. Establishing the validity of fewer decision points and proficiency levels (See section 13 for a full definition of this term.) is more readily accomplished than if more points are used. MSDE should consider using a fairly small number of cut points (See section 13, for a full definition of this term) at which decisions will be made. As noted above, development of a substantial number of items that measure at a decision point is essential to effective discrimination between students who perform above and below that point. Supporting two or three decision points with a single test will require more items at those points. The consequences of this are: (a) to require more total items on the test, longer tests, and proportionately more SR items to provide decision validity at each point, and (b) to result in fewer items along the entire distribution of student performance which will significantly reduce the utility of the test for purposes of remediation. The same argument applies as the number of proficiency levels increases. If MSDE decides to establish five proficiency levels with four

decision points for each test, it is highly likely that these tests will either have to contain substantially more SR items than proposed or the tests will have less validity and utility for remedial purposes.

Determination of the level of decision validity of an assessment instrument requires the gathering of data over time. The levels into which students have been classified as a result of test scores must be compared to their levels of performance on tasks for which their high school education served as preparation. Collection of the latter performance data is, unfortunately costly and time consuming.

Recommendation: CB/ETS recommend that the study of decision validity be a research activity carried out under MSDE direction.

3. Reliability and Numbers of Items

Reliability refers to the precision with which a test assesses the knowledge, skills, and other components which it is designed to measure. The importance of decisions to be based on HSA results dictates that reliability be very high. Reliability coefficients range from 0.0 (no precision whatsoever) to 1.0 (perfect precision). Highly accurate tests are, necessarily, comprised of sufficient numbers of precise items. The precision of a test item is measured by *item discrimination*, the degree of relationship between the item and the dimension assessed by the test. In general, the larger the number of items on the test the higher the reliability. This relationship is, however, determined by the types of items and their statistical characteristics, such as discrimination. Besides item discrimination, the distribution of *item difficulty indices* of a test form is an important aspect of test reliability.

As in the case of validity, reliability must be considered in light of uses to be made of assessment instruments. Accurate placement of students along a dimension measured by a test requires that all test forms meet the same very high reliability requirements. In addition, the use of the test to classify examinees into levels of achievement requires that the classification of students be precise. Thus, very high precision is required at the points along the scale that serve as cutting points to classify individuals. Finally, identification of remedial needs requires precise measurement of elements of the test content relating to that diagnosis.

3.1 Overall Target Reliability

Because of the importance of decisions to be made using results from the HSA instruments, overall reliability coefficients for each test must reach or exceed the .90 level. This specified level refers to internal consistency indices of reliability, such as *coefficient alpha*, that may be estimated from the results of a single administration of the test. To reach this criterion will require large numbers of selected-response (SR) items, also known as multiple-choice (MC) items. Use of SR items is an efficient way to cover a broad content domain because larger numbers of these items may be administered within specified time constraints than other item types such as constructed response (CR) items. There will be a

disagreement between the desire of educators to have a substantial proportion of testing time (e.g., 50%) devoted to CR items and the psychometric requirement for a target reliability of .90 that would be needed to defend high-stakes uses. Initial field testing should be conducted by MSDE prior to the development of complete testing forms to estimate the reliability of the HSA given the current specifications and determine if additional modifications are required to meet the target reliability within each test.

3.2 Reliability of Multiple Decision Points

If the sole purpose of a test is to classify students into two levels such as pass/fail, then highly precise measurement is required at a single point in the dimension being assessed. In this use of a test, there is no need to differentiate between different degrees of pass or different degrees of fail. The test should then be comprised of a large number of items that precisely measure (have high discrimination indices) at that single point. Such a test would have a narrow range of item difficulty indices concentrated at the *cut point*, the scale point where the differentiation is made between pass and fail.

If the test is to be used to classify students into more than two classes, then there are multiple cut points and it is important that the distribution of item difficulty indices reflect that usage. In other words, there should be concentrations of highly discriminating items at the multiple cut points.

3.3 Reliability for Identification of Remedial Needs

In contrast to the above, remedial needs may require that the distribution of test item difficulty indices be uniform over a wide range of scale points. The reason is that it may be important to locate each student very precisely along the dimension being assessed. Remedial measures for a student nearly reaching the pass/fail cut point may be quite different from those for a student scoring at a slightly lower level, and different again from those for a student falling far below the cut point. Decisions about formal courses of study would differ for students who pass at different levels. Given that the population of students would be expected to be distributed over the entire range of scale points, this usage requires precise measurement over the entire score range. This requirement is clearly in disagreement with the need to concentrate precise measurement at a few cut points (See section 13 for a full definition of this term) in the scale. Tests that are most precise for high-stakes decisions may not be effective for remedial purposes because of the concentration of items at the decision points, rather than across the range of achievement levels. The same test cannot be equally effective for multiple purposes or uses. Maryland constituents may desire that the tests be used for multiple purposes, but this is simply not a reality.

Maryland HSA tests will provide subscores (See section 13 for a full definition of this term) for each of the twelve tests. This subscore information is intended to provide some general indication of student performance in several areas. If precise diagnostic

information is desired, more stringent psychometric standards would apply to the development and selection of items for each subscore.

Recommendation: CB/ETS recommend that MSBE prioritize the proposed uses so tests can be designed to support the most essential uses with the appropriate level of reliability and validity, and once this is attained, attention can be shifted to supporting secondary uses.

3.4 Overall Reliability Considerations

Given the several purposes of the Maryland High School Assessment, it is clear that a large number of highly discriminating items with a wide distribution of item difficulty indices is required. For most of the HSA tests, this will dictate a large ratio of SR to CR item types. Examination of data from other assessment programs indicates that this ratio depends on the subject matter (generally mathematics and science do not require as large a ratio as do English or social studies) and types of items used. Some guidelines based on study of other instruments used for assessment at the high school level are provided in Section 14.

3.5 Recommended Numbers of Items

Based on information from the College Board Advanced Placement (AP) Program (Section 14), it must be assumed that the level of reliability that can be attained with various mixes of item types is highly subject-area specific. That is, the ratio of SR to CR items differs markedly across subjects. This section contains a summary of the information provided in Section 14.

3.5.1 English

In the AP programs in English, 55 to 65 SR items did not yield reliability coefficients as high as the Maryland HSA criterion of .90. And when CR items are added such that they accounted for 60% of the score points, the reliability was lower for the composite than for the SR items alone. It would take about 100 similar SR items to achieve a reliability of .9 and adding a few CR items could reduce that figure. This compares to the specification committee recommendation of 90 SR items in English.

3.5.2 Mathematics

AP mathematics displayed a different picture. Based on data in the Appendix, CB/ETS estimate that a mathematics test comprised solely of about 60 SR items like those on the AP tests would yield a reliability coefficient in the .90 range. And in no case of the AP examples did the addition of CR items decrease the reliability. The mathematics specification committee recommendation was 70 SR items.

3.5.3 Science

The AP science tests are somewhat similar to those in mathematics. Although there is some variation from one science discipline to another, for items like those on the AP tests it would appear that about 70 SR items plus some smaller number of CR items would result in the required level of reliability. The number of science items recommended by the specification committee was 90.

3.5.4 Social Studies

On the American and European History tests of the AP program 100 SR items yielded a reliability coefficient of about .90. Adding CR items to the extent of one-half of the score points decreased this figure substantially in many cases. In comparison, the specification committee recommended 60 SR items in Social Studies.

3.6 Allowing for Field Test Items

If items are to be field tested on operational forms of the Maryland HSA, as discussed below, allowance must be made within the testing time limits for additional field-test SR items to be administered. This might result in fewer CR items than desirable because, to reduce the number of operational SR items to allow for the field test items could reduce the reliability below the acceptable level of .90. Another option is to accept a lower level of reliability by reducing reliance on the HSA for high-stakes individual uses in order to support other proposed uses (e.g., stimulate changes in instructional practices, reform education, support Core Learning Goals).

4. Distributions of Items by Difficulty Levels

In order to specify distributions of items by difficulty levels, it will be necessary to take into consideration decisions with respect to some of the reliability and validity issues discussed above. As was pointed out above, if it is important to have the same degree of precision at all scale points, then the items' difficulties should be uniformly distributed over the score range. If, on the other hand, certain scale points will be used for high-stakes decisions, it is important to measure more precisely at these points, which will result in a greater proportion of items having difficulty levels at selected points rather than other points in the scale.

The problem of specifying appropriate distributions of item difficulty indices is made more difficult by the fact that it is hard, if not impossible, for test developers to know the exact difficulty levels of all items before any data have been collected from students' interactions with the items.

Recommendation: For these reasons, CB/ETS recommend that HSA test developers:

- develop items that appear to be distributed fairly uniformly across the entire range of the scale
- develop many more items than will be required on the final forms of the tests
- do an extensive field testing of those items (may be the developmental phase of the HSA)
- study estimates of item statistics from the field tests
- determine the appropriate distribution of difficulty indices according to the purposes of assessment, discussed above
- select items for each form such that each form has the appropriate distribution of items

5. Field Testing, Calibrating, and Scaling Items

In order to construct test instruments that provide valid and reliable measurement, thorough field testing of items is needed. Many programs field test two to three times as many items as needed for the final (operational) test forms.

Recommendation: CB/ETS recommend that at least twice as many items be developed as would be needed for operational tests.

This number, which applies to both constructed-response and selected-response questions, is based on comparable Advanced Placement item performance information. Harvest rates (proportion of items that survive field testing and are used on operational test forms relative to the total number of items developed for field testing) for individual content areas may vary from this overall number. For example, it may be necessary to field test three times as many English questions as needed, but only 1.5 times as many Mathematics questions. No-fault data can help to inform future item development requirements.

In general, field testing takes one of two forms: special field test samples or field testing new items on operational forms. Many testing programs use both, field testing initially on small samples and then inserting the items on operational test forms. The initial field test provides information used to refine the items before they are administered to larger samples of examinees. If field testing is done on operational forms, the field test items are not used in determining examinees' scores. This type of field testing is efficient in achieving accurate estimates of item statistics from large numbers of student responses without requiring special sampling, administration, etc. Accurate estimation of item statistics is a requirement in order that test forms meet the needs of the Maryland HSA mentioned above.

Recommendation: CB/ETS recommend that selected-response questions be field tested on operational test forms and that constructed-response questions be field tested on special samples recruited from out-of-state populations.

This latter recommendation will ensure students from some schools in the state do not have early exposure (and an unfair advantage) to CR items that will be used on operational test forms.

5.1 Selected Response Items

It is recommended that SR items be field tested on operational forms of the HSA in order to achieve the efficiency mentioned above. The no-fault administrations can be used to field test a large number of items for the initial test forms. In this case, these administrations will serve as a developmental assessment, with the data being used to estimate item statistics and not to report individual scores. Once the HSA tests are operational, it will be necessary to develop an ongoing source of items to replenish the item pools for security reasons. It is recommended that SR items be pretested within the operational forms, with each student receiving only a small number of these items. The exact number of items to be pretested will depend both on the number of students taking a test and on the number of printer spirals (and printer costs) to be incurred. The number of SR items taken by any individual student should be minimized.

Recommendation: CB/ETS recommend that 5 - 10 selected response items be included on each examinee's test to pretest the questions. These questions would not count toward an examinee's score, but would be used to help build a pool of test questions from which subsequent tests can be assembled. This low number should be possible on most of the tests, with the possible exceptions of Chemistry and Physics where very low examinee volumes might necessitate higher numbers of pretest items.

5.2 Constructed Response Items

The above-mentioned efficiency cannot truly be attained with the large numbers of CR items that will need to be field tested. Because it is not possible to pretest CR items within the operational test (this would consume too great a proportion of total testing time), it will be necessary to field test these items in special samples where students are exposed to content similar to that prescribed by the Maryland CLGs. It will clearly be crucial to the accuracy of estimation of item statistics, and the assembly of items into reliable and valid tests, to ensure that the field test administrations are done under conditions like those of the operational HSA. The conditions of importance include using the same directions and timing in test administration, and presenting the field tests to the students in such a fashion that they are as highly motivated to do their best as they will be under operational conditions. It will also be important to ensure that the field test sample is representative of the total test group so that accurate statistics can be gathered.

Recommendation: CB/ETS recommend that CR field test data be gathered on out-of-state samples of examinees in order to maintain the security of the CR questions before administration.

5.3 Item Response Theory (IRT) Calibration and Scaling

Many testing programs use IRT calibration and scaling of assessment instruments. Calibration refers to estimation of item parameters from data resulting from the interaction of samples of the target population with the items. Scaling refers to the methodology used to determine examinees' scale scores (See section 13 for a full definition of this term.). Any type of scaling that requires precise determination of those scores requires precise estimation of item parameters. This, in turn, requires large representative samples of examinee-item interactions for calibration. Inclusion of CR items that yield multi-point (polytomous) scores, as compared to SR items yielding dichotomous (right/wrong) scores, requires larger samples of students in field tests. For tests comprised strictly of SR items, a sample size of 1000 would be desirable. For polytomously-scored CR items, especially those involving highly subjective judgments in the scoring process, larger samples are required. In the National Assessment of Educational Progress, for example, the specified sample size for item calibration is 2000 students.

5.4 Other Scaling Alternatives

Because IRT scaling results in an arbitrary scale, some testing programs prefer to use a simpler procedure such as raw score scaling. In the latter approach, the scores of each student on each item are simply summed to yield an overall test score. For instruments comprised solely of SR items, this is often referred to as "number-correct scoring." When this type of scaling is used, it is imperative that each test conforms very precisely to a set of statistical specifications. Otherwise, the test forms may not be equivalent and the scores assigned to a student may be highly dependent on the particular form of the test that was administered.

The main advantage of IRT scaling is that, once all test items are precisely calibrated, students tested with different sets of items may be placed precisely on the same scale. It must be emphasized, however, that this requires precise calibration of items that meet the assumptions of the IRT models employed. The main assumptions are unidimensionality of the test items (all items are measuring the same single dimension in the content area) and conditional independence of items¹. The first assumption can be summarized as requiring that each student interacts with each item using the same combination of skills and abilities, and that all students use this same combination. The latter assumption means that for students at the same point along the dimension being measured (the "conditional" part),

¹ Although the assumptions of IRT calibration and scaling may appear to be more restrictive than those of simpler scaling procedures, the same assumptions truly are necessary for precise interpretation of scores based on alternatives. For any scaling, items that are not unidimensional will yield the same scores from students interacting with the items using different combinations of skills and abilities. Hence the same score may mean different things for different examinees. Also it has been pointed out that lack of conditional independence results in an overestimate of test reliability. Hence, although these assumptions are usually explicitly stated only for IRT scaling, they are implicit in interpretation of other types of scores.

their responses to each item are independent of their responses to every other item on the test.

5.5 Calibration Sample Sizes

Based on considerations discussed above, it is recommended that a *minimum* of 1000 students be used in calibration of SR items and 1500 students for CR items. These numbers should yield reliable estimates of item parameters.

There has been some discussion that certain students (e.g., Advanced Placement students) might not be required to take the HSA tests. Such a decision may have implications for calibration samples. If the population for which the tests are designed excludes these students, and the instruments are constructed and field tested with this taken into consideration, there may not be any problem. If, however, the intention is to define the populations of interest as all students at a given grade, excluding part of the population from samples would cause a truncation in scores and a decrease in variance. This in turn could result in lower reliability estimates and biased item statistic estimates (if those students are included in calibration samples but not in the operational administrations of tests, or vice-versa). Similar differences could be encountered if students from private schools are later added to the test taking population. MSBE should first define the population of students who will be required and eligible to take the HSA (e.g., transfer student rules, private school students on a voluntary basis, home schooled students, students with disabilities), and this should occur prior to field testing. Next, field testing should be based on a representative sample of students. Geographic region, ethnicity, social economic indicators, student accommodations, and achievement levels of the sample relative to the intended population should be considered in constructing sampling designs for the field testing.

6. Equating

Equating refers to statistical procedures used to ensure that examinees' scores derived from different forms of a test are on the same scale. Assuming that the different test forms precisely follow content and statistical specifications, they may still yield scores on slightly different scales. Equating "lines up" the scales so that scores are consistently reported on the same scale. For properly scaled and equated multiple-form assessment instruments, it will not matter to any given examinee which form is used for a given administration. Within measurement error, the results will be the same.

6.1 IRT Equating

For initial forms of a test, as mentioned above, items must be administered to a large field test sample of examinees. It is not necessary that all students respond to all items. A matrix-sampling scheme such as that used in the National Assessment of Educational Progress NAEP (Allen, Kline, & Zelenak, 1996) may be used. IRT methodology similar to

that used in NAEP can be used to calibrate all the items, and then the required test forms can be assembled using the content and statistical specifications.

When IRT methodology is used in conjunction with field testing on operational forms of the assessment, equating is relatively straightforward. Joint analysis of the operational and field test items can be used to calibrate the new items on the same scale as the already-calibrated operational items. Following calibration of the new items, new forms can be assembled by following the content and statistical specifications for the test.

6.2 Other Forms of Equating

If IRT methodology is not used, the test forms must be equated through either common population or common item equating. In the former, different randomly equivalent subsamples of a population are administered the different forms. In common item equating of two forms, two samples of examinees are administered a common core of items along with the different items from the two forms. In either case, standard equating methodology (Kolen & Brennan, 1995) may then be used to ensure that scores from the different forms are on the same scale.

7. Scoring Standards

7.1 General Considerations

In the case of SR items, machine scoring should be used and procedures should include certain quality control checks of the resulting data files.

7.2 Treatment of Missing Item Response Data

Typically, some examinees will not respond to all items on a test form. Many assessment programs differentiate between two types of nonresponse; labeled *not reached* and *omits*. Decisions must be made about how missing item response data will be treated in the scoring process.

7.2.1 Not reached Items

Some examinees will produce answer sheets that have a string of missing responses at the end of the test. It is usually assumed that if the student's last response is to the *n*th item that she/he did not have (or take) the opportunity to interact with the remainder of the items. The items following the *n*th are labeled not reached for that student. In deriving a score for the student, a decision must be made about whether the not reached items should be treated like incorrect responses or like items that were never administered to the student. Current practice in assessment programs such as NAEP is to do the latter. In this case the student's score is not affected by the not reached items.

Recommendation: CB/ETS recommend that HSA follow current practice and treat the not reached items as if they had not been presented rather than as incorrect responses.

7.2.2 Omitted Items

Some examinees will not produce a response to an item but will respond to items that follow it on the test. This type of missing data is referred to as an omit. The most common assumption about omits is that the student had an interaction with the item but was unable to respond. Under this assumption, the response is treated like an incorrect response in the scoring process. In this case the student's score is affected by the missing response data.

Recommendation: CB/ETS recommend that omitted items on HSA tests be treated as incorrect responses.

7.3 Scoring Constructed-response Items

In the case of CR items that must be hand scored by human scorers, the question of reliability of the scoring procedures is of concern. A student's score on a test should not be highly dependent on the individual doing the scoring nor the time at which scoring is performed. The same scorer rescoring a student's paper a second time may not assign the same score to an item. And different scorers may also assign different scores to the same paper. To the extent that item scores vary between scorers or within scorers over time, there is unreliability in the scoring process. Although various kinds of scoring differences can arise, procedures such as those discussed below can be used to lessen the chances that these effects will occur or that they will affect results in a significant way.

Certain procedures are used to maximize scorer reliability in testing programs. One overarching principle is to standardize the scoring process. Standardization begins with very specific scoring directions (rubrics) for each item and includes extensive training of scorers in use of the rubrics for each item and in the rating of sample responses. Typically, rescoring a certain percentage of student responses (e.g., 10%) is done and the reliability of the scoring is thereby estimated. For high-stakes testing programs, two or more scorers are required to ensure inter-rater reliability. Some information on scorer reliability is presented in Section 14.

Recommendation: CB/ETS recommend that the HSA tests incorporate multiple scoring of a sample of student protocols on each CR item and that interscorer reliability be estimated. One simple statistic that may be used is the percentage of exact scorer agreement (both scorers give the same item score to the paper) for each item. Another statistic is the correlation between the scores assigned by two raters.

As might be expected the degree of subjectivity in the rubric is related to the level of reliability that can be attained. In content areas such as mathematics problem solving, the

rubrics might be very specific and the scoring may not be subjective at all. In other content areas, particularly for cases in which essays or shorter writing exercises are necessary parts of the content specifications, there is more subjectivity. Scorer agreement will be lower for such test items.

Recommendation: CB/ETS recommend that MSDE use a generic scoring rubric for each test's extended constructed response item; this scoring rubric could be made public so that teachers, parents, students, and others understand the criteria on which the student work will be evaluated. The generic rubric could be supplemented, if needed, with additional scoring instructions for a particular item. For both the brief constructed response (BCR) and extended constructed response (ECR) items, CB/ETS recommend that a limited scoring range be used. A scale of less than six points should be considered. Depending on the BCR items that are generated, a more truncated BCR scale might be appropriate. If a truncated BCR scale is used (e.g., 2 - 3 score points), it may be possible to achieve acceptable scoring with a single rater. Investigations into this approach could be made during the no-fault administration. Additional discussions on these issues are contained in the interim report.

8. Quality Assurance (QA) Checking of Data

Errors can occur in creating computer data files from the student papers and scoring protocols. Many assessment programs, including NAEP (Allen, Kline, and Zelenak, 1996), specify an extensive quality assurance checking of data prior to scaling and other analyses. Typically, this involves taking a small sample of student test papers and scoring results and comparing them with electronic data files to determine the error rate.

A second QA procedure that should always be used is analysis of data in the data files to check that the results are within expected ranges. Often an initial item analysis will be conducted on a subset of data that is available early in the process of creation of data files. It may be decided, for example, that whenever 10% of the data are in electronic form, an item analysis will be carried out and closely scrutinized for anomalies. With a sufficient sized sample, item statistics in this preliminary item analysis should closely approximate those from the earlier calibration samples. Such a QA procedures may also be useful in detecting possible compromising of items (discussed further below).

The results of all analyses during the entire process should, of course, also be closely studied to verify that the results are reasonably accurate.

9. Post-administration Analyses

A number of standard analyses should be conducted when all the data have been entered into computer files.

9.1 Item Analyses (IA)

For each assessment a standard item analysis should be conducted. A typical IA will report, for each item:

- a difficulty index such as the proportion of correct responses on MC items and average score on CR items
- a discrimination index such as the biserial (dichotomously scored items) or polyserial (polytomously-scored items) correlation between the item score and the total test score (usually the raw score rather than a scale score)
- the frequencies of responses to each alternative of a MC item and to each possible score of a CR item, and the frequencies of omits and not reached

An IA containing at least the information outlined above should be performed on each test form used in each HSA assessment. In addition, as mentioned above, early IAs on a certain percentage of the data should be performed because this procedure may help to detect problems in the scoring and data entry procedures.

9.2 Differential Item Functioning (DIF) Procedures

It is common practice to examine all test items for potential inequity between various subpopulations of examinees, usually defined by gender and majority/minority racial-ethnic groups. A basic principle of good measurement is that performance on each item should reflect the examinees' abilities and no other factors such as membership in a particular subgroup of the population. Statistical procedures referred to as DIF procedures provide information about whether students of differing subpopulations who have the same ability level have different performance on the item. These procedures may identify items on which the subgroups really differ in their level of knowledge as well as items that actually reflect an aspect of inequity. Hence the statistical procedures, by themselves, cannot determine definite presence of inequities. Therefore, based on statistical DIF criteria, rules should be specified for flagging items that produce different results for different subgroups while controlling for ability level. Each flagged item should be closely examined by experts to determine whether the item may be biased against a particular subgroup. The set of experts should include individuals who are knowledgeable about measurement as well as those knowledgeable about aspects of item content (e.g., language) that may induce inequity for different groups of individuals.

Recommendation: CB/ETS recommend that a statistical DIF procedure be conducted on the items of each new test form. As a first approach to utilizing DIF information, CB/ETS recommend that all items yielding extreme DIF statistics (i.e., those which evidence large group differences in favor of either examinee group being compared) be eliminated from the HSA item pools. This procedure should be carried out with respect to gender groups and with respect to majority/minority racial/ethnic groups. Because DIF procedures do not

provide reliable information for small subgroups they should not be used with subgroups of less than 200 students.

The amount of the difference for items to be flagged as extreme should be determined by MSDE staff, based on their review of the no-fault data. After the first year of operational administration, MSDE staff may wish to set a more restrictive range for the DIF values or they may wish to set a mean DIF value for the tests. If DIF data for all subgroups of concern are not available at the pretest stage, then it is recommended that a panel of reviewers be convened after DIF analyses on operational data but before scores are generated. The purpose is to determine whether any item(s) should be dropped from scoring because of possible inequity.

9.3 Speededness Analyses

Most assessment instruments are designed to be power rather than speeded tests. By “power” we mean that most examinees should be able to finish the test in the time allotted without being rushed. The assumption is that the best assessment of individuals’ levels will be accomplished if they have had the opportunity to interact relatively unhurriedly with all test items. To determine the degree to which a test is speeded, the proportion of examinees responding to each item is examined. For a test to be nonspeeded, some large proportion (e.g., .80) should have interacted with the last item of a test. In other words the not reached rate should be small for each item on each final operational form. This indicator of test speededness is obviously an estimate of the percent of students completing the total test (not including items omitted). That is, a relatively high proportion of students (e.g., 75-85%) may be expected to complete the test if a very low level of speededness is desired.

9.4 Analysis for Detection of Cheating

Certain analyses may be conducted to assess whether tests or test items have been compromised. The HSA should determine whether or when such analyses should be conducted.

One analysis that is often used is to check for copying through study of the response patterns of examinees who were seated in adjacent seats during the test. The incidence of identical responses can be compared to probabilities of getting identical response patterns by chance. Such an analysis relies on the availability of accurate seating charts, which contain information about the test booklet used by each student.

Another commonly used procedure is to study item difficulty levels over consecutive administrations of each item. If the item is becoming less difficult, the reason may be that the test items are no longer secure. This analysis is particularly necessary if test items will be used on more than one form in the same or consecutive years. CR items, because of the length of time students typically interact with them, are more likely to be remembered than SR items. Thus they present the biggest challenge and are not reused in many national testing programs with high-stakes consequences.

Recommendation: CT/ETS recommend that HSA not reuse CR items unless the reuse is at irregular intervals with several years in between.

10. Reporting Procedures

The method of score reporting must be determined. Among the decisions to be made is whether scores will be reported as point estimates or as point and interval estimates. The latter would be based on estimates of the standard error of the test. One problem with interval estimates is that the interval will, for many students, include cut points between proficiency levels. The resulting ambiguity of level may present problems in high stakes assessments.

Recommendation: CB/ETS recommend reporting point estimates in order that ambiguities of classification into levels be avoided.

11. Weighting Issues in Combining Items or Subtests

An important issue in constructing tests that use both multiple-choice and constructed-response questions is “How should scores on individual items or sets of items² (subtests) be combined to form an appropriate total test score?” In addressing this issue, most individuals involved in test design think in terms of the contribution of items (or subtests) to the total test score (or test score variation). However, the answer to the question is not an easy one for several reasons (Carlson, 1968; Gulliksen, 1987, Chapter 20; Wang and Stanley, 1970). There is no unambiguous definition of contribution except in the case of zero correlations among all items (Carlson, 1968). That case, of course, would represent poor test construction in that each item would be measuring a distinct skill or trait that is uncorrelated with those measured by all other items.

11.1 Intuitive Weighting

The following quote from Gulliksen (1987) correctly specifies the fact that the kinds of intuitive weights (Wang & Stanley’s, 1970, “nominal weights”) often used in test assembly are essentially irrelevant to the “contribution” of items or subtests to the total score.

In most amateur discussions of weighting of tests the first factors considered are the number of items in the test and the average magnitude of the score. It is believed, for example, that if gross scores are added, the effect will be to give a 100-item test twice the weight of a 50-item test. That such is not the case can be seen for example by assuming that the 100-item test was a very easy one on which everyone obtained scores ranging from 95 to 100. Adding scores on this test to a student’s record would then, at the most, make a

² Any set of items may be arbitrarily defined as a subtest even though there is no interest in reporting scores on those “subtests.” Hence, when a total test score is formed from N multiple-choice items and M constructed-response items, the two sets of items can be considered two subtests of sizes N and M. Ultimately, because each item could be considered as a subtest of size 1, combining individual items is similar to combining subtests.

5-point difference in the total score. If, on the other hand, the 50-item test were composed of fairly difficult items and were fairly reliable, it could easily be that scores on it would range from 20 to 50. In other words, adding this test would make a 30-point difference in extreme cases, and a 10- or 20-point difference in the majority of cases, so that the total score would agree rather closely with the score on the 50-item test and not correlate with the score on the 100-item test (Gulliksen, 1987, pp. 36-37).

11.2 Contributions to test score variation

Following through on Gulliksen's point, it is clear that it is the spread of scores on each subtest that is relevant to the spread of scores on the total. Spread, or variation, in scores is measured by the variance, or its square root the standard deviation. Unlike variance, which is in the metric of squared scores, the standard deviation is in the same metric as the scores. For that reason, the standard deviation is somewhat easier to interpret than the variance.

The standard deviation of a composite test, C, made up of two subtests, X and Y, is a function of the variances of the two subtests and the correlation between them. The formula for the relationship is

$$\sigma_c^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \quad (1)$$

where, σ_c^2 , for example, represents a variance and σ_{XY} a covariance between X and Y. The covariance is equal to the product of the two standard deviations and the correlation between the two variables. It is clear from equation 1 that if the two subtests are standardized to zero mean and unit variance before forming the composite, the variance of the composite, Z, will be simply

$$\sigma_Z^2 = 1 + 1 + 2\rho_{XY} \quad (2)$$

Note that the standardized covariance is the correlation coefficient, ρ_{XY} . Further, if after standardizing, the X subtest scores are weighted by a constant, k, the weighted variable's standard deviation will be k, its variance k^2 , and the variance of the composite, W, will be

$$\sigma_W^2 = k^2 + 1 + 2k\rho_{XY} \quad (3)$$

The last term of equation 3 reflects the fact that if both X and Y are standardized and then the standardized X is weighted by k, the covariance is $2k\rho_{XY}$.

As an example, suppose a set of multiple choice items (subtest X) is combined with a set of constructed response items (subtest Y), standardizing before forming the composite and weighting X by $k=3$. Suppose also that the standard deviations of X and Y are both 10, and

the correlation coefficient is .7 (hence the covariance is 70). Then it is true that:

$$\begin{aligned}\sigma_C^2 &= 100 + 100 + (2 \times 70) = 340 \\ \sigma_Z^2 &= 1 + 1 + (2 \times .7) = 3.4 \\ \sigma_W^2 &= 3^2 + 1 + (2 \times 3 \times .7) = 14.2\end{aligned}$$

Wang and Stanley define variance contributions as “effective weights” and suggest defining these for X and Y (using equation 1) as $\sigma_X^2 + \sigma_{XY}$, and $\sigma_Y^2 + \sigma_{XY}$, respectively. Most psychometricians would agree that the two subtests are contributing equally to the composite variation in the case of the C and Z composites. And using Wang and Stanley’s definition, both effective weights would be 170 for composite C and 1.7 for composite Z. But what are the individual contributions in the case of the W composite? Wang and Stanley’s definition would indicate that they are 11.1 for X and 3.1 for Y. In their terminology, k, above, would be referred to as a nominal weight.

In this example most test developers would assume that it is the number of items and the nominal weights that determine “contribution.” Such a definition of contribution involves “naïve” weights. It is quite possible that the example statistics cited above resulted from X made up of 20 multiple-choice items and Y of 5 constructed-response items. The naïve weights, after multiplying X scores by 3, would be $20 \times 3 = 60$ for X and 5 for Y. Another approach is to assume the total possible score defines weighting. Hence, if the example constructed-response items each yielded 5 points in the scoring, the naïve weights after weighting would be 60 for X and $5 \times 5 = 25$ for Y. Another assumption made by some test developers is that the amount of time devoted to each subset of items indicates the relative weighting. In the example that assumption depends on what assumptions are made about the relative amounts of time spent on the two sets of items.

This example well illustrates the ambiguity associated with the notion of contribution of item to total test score. Also, using the notion of effective weights, it is clear that one needs to know all of the subtest parameters (variances and covariances) before the effective weights can be determined. Although it might be argued that effective weights may vary over time while the “nominal” weights do not, the numbers being thought of as nominal weights are not truly weights but factors that affect the weights (hence we refer to them as nominal weighting factors below). Item parameters varying over time (“parameter drift”) is a concern not only for weighting but also for consistency of scores from the assessment instruments. Variation in item parameters changes the scale!

11.3 Weights of Individual Items

A further consideration is that each item may be thought of as making an individual contribution to a test. Using the Wang and Stanley definitions, the effective weight of an item (without use of nominal weighting factors) is the item variance plus the sum of the covariances of that item with each of the other items on the test. Note that, in the case of dichotomously scored items such as the SR type, the item variance is the product of the p-

value (proportion correct) and 1.0 minus the p-value. Hence, in that case, the effective weight is directly related to the item difficulty level. In the case of polytomously-scored items such as the CR type the effective weights depend on the variation in scores on the individual items.

As in the case of effective weights for subtests, it is necessary to know all of the item parameters before effective weights of items can be determined. And in this case we are not discussing classical test theory parameters but item standard deviations and covariances. It is true, of course, that the classical test theory parameters are functions of the standard deviations and covariances as is pointed out above and in Gulliksen's book.

The alternative of using IRT scaling to determine weights suffers from the same problem as effective weights – one needs to know the item parameters to determine the contributions (which are roughly proportional to the a-parameters).

Wang and Stanley (1970) point out that "With a large number of positively correlated variables (such as test items), the correlation between two randomly weighted composites rapidly approaches unity (p. 699)". Hence it is probably best in most assessments to let the items weight themselves according to their parameters rather than to use nominal weighting factors.

11.4 Conclusions

This discussion of weighting issues is meant to show that the question of effective weighting of various components assembled into a test form is a very complex issue with no easy solution. Decisions about weighting should be made only after careful consideration of no-fault data, which will yield item parameter estimates.

Recommendation: CB/ETS recommend that the HSA not use nominal weighting factors.

12. Reuse of Test Questions

When important decisions about individuals are being made, it is crucial that examinees not be able to anticipate the questions that they will be asked. If the same test questions are administered on different test dates, it becomes easier for examinees to achieve a high score without knowing the domain of knowledge. The problem of maintaining security of test questions is especially serious for constructed-response questions. Since these questions are usually very easy to remember, they are difficult to keep secure. For this reason, the Advanced Placement program does not reuse any constructed-response questions. On a high-stakes test, like the HSA tests, it is very risky to re-use constructed response questions in the regularly scheduled examinations, even after a period of time. Even if students do not keep files of previously used questions, teachers surely will.

Recommendation: CB/ETS recommend that HSA constructed-response questions be used only once in a regularly scheduled test and once, after an extended period, in a make-up test.

Selected-response questions are, by their number and nature, much more difficult to remember. As pointed out above, students generally have longer interactions with CR items and are therefore more likely to remember details about them. Thus, it is possible to reuse SR questions more frequently than CR questions. However, there should not be over-exposure of these questions, and items should be reused only a small number of times. For example, in the Advanced Placement program, selected-response questions are used a maximum of three times.

In a high stakes assessment like the HSA it is highly likely that groups of students purposely plan to compromise the tests by memorizing a few items each during the test administration. They may assume that several of them will have to retake the test. Collective memorization by a number of students can quickly compromise an entire test form if it is used repeatedly.

Recommendation: CB/ETS recommend that the HSA program limit the number of times selected-response items can be used to three or four. To the extent possible, large numbers of selected-response items should not be reused in adjacent test administrations.

13. Glossary of Terms

Cut scores

A ***cut score (cut point)*** is a point in the score scale that is the dividing lines between two performance level categories. When students are being assigned letter grades based on a total score for the work in a term, for example, a teacher may decide that a total score of above 84.5 is an A and a total score below that value is a B. In this example 84.5 is the cut score.

Multiple levels of performance

The term ***multiple levels of performance*** refers to a set of ***proficiency levels*** (see below) based on scores on an assessment instrument. The multiple levels are typically described by terms such as “failing”, “passing”, “passing with distinction”, “honors”, etc.

Proficiency levels

A ***proficiency level (achievement level)*** is a category of performance that is used to classify students. A simple example of the use of proficiency levels is letter grades used by a teacher for reporting students' classroom work. Proficiency levels are operationally defined by ***cut scores*** on the assessment instrument used to measure performance. Five proficiency levels, for example, would be operationally defined by setting four cut points in the score scale.

Scale scores

A ***scale score*** is a score on the ***reporting scale*** of an assessment instrument. In assessment programs it is necessary to report scores in a metric that will be clearly understood by the intended users (e.g., students, parents, teachers, school administrators). Furthermore, different test forms must be scaled to the same metric (through psychometric equating) in order that scores of students responding to the different forms be comparable. If the different instruments in a program are comprised of different numbers of items, or of items having different average difficulty, a simple "number of items correct score" does not provide precise information to the user about a student's performance relative to standards or to other students. Scaling refers to the psychometric procedure of transforming the scores from a given assessment instrument to the scale scores.

Subscores

A ***subscore*** is a score on a subset of the content domain of an assessment instrument. For example, a mathematics test may be designed to yield a ***total score*** and subscores in the subdomains: numbers and operations, data analysis and probability, geometry, algebra, and calculus. Subscores may be precise indicators of performance on a defined set of content or skills built to strict psychometric standards of difficulty, or as in Maryland, they may be more general indicators of performance.

Section 14.

In this section are some examples from existing programs of reliability levels and other statistics that have been attained with certain item numbers and combinations of item types. Most of the results are from the technical manual of the College Board's Advanced

Effects of Scoring on Test Reliability: College Board Advanced Placement (AP) Tests

The following table presents scorer reliability coefficients that are reported in the technical manual. These coefficients are estimates of test score reliability of tests comprised of a mix of MC and CR items. Separate estimates are reported for single reading of each CR item and for double readings.

Table A1. Effects of Scorers on Test Reliability

Subject Area	Year	No. Items MC/CR	Type of Reading	Reliability Estimate
English Lit. & Composition	1984	58/3	single	.84
			double	.91
	1982	55/3	single	.84
			double	.89
English Lang. & Composition	1984	55/3	single	.82
			double	.86
	1982	60/3	single	.84
			double	.90
American History	1981	84/1	single	.79
			double	.90
European History	1981	90/1	single	.78
			double	.88
Chemistry	1980	80/6	single	.95
			double	.97
Physics B	1981	70/6	single	.96
			double	.97
Physics C, Mechanics	1981	35/3	single	.94
			double	.94
Physics C, Elect. & Magnetism	1981	35/3	single	.95
			double	.95

Test Reliability: College Board AP Tests, Mixes of Item Types

In this section are presented reliability estimates for various mixes of item types on certain AP tests. Data were reported by maximum possible score and percentage in the composite rather than number of items of each type.

Table A2. AP Reliability estimates by Item Types

Test	Year	Max Possible Score MC/CR/Comp. (%CR in Comp)	MC Reliability	CR Reliability	Composite Reliability
English Language and Composition	1982	60/100/150 (67)	.85	.57-.78	.78-.87
	1983	55/90/150 (60)	.76	.62-.75	.78-.82
	1984	65/90/150 (60)	.85	.58-.82	.82-.89
	1985	65/90/150 (60)	.88	.56-.82	.82-.90
	1986	60/90/150 (60)	.84	.60-.82	.80-.88
English Literature and Composition	1982	55/100/150 (67)	.86	.57-.75	.79-.86
	1983	58/90/150 (60)	.81	.61-.75	.80-.85
	1984	60/90/150 (60)	.82	.55-.75	.79-.85
	1985	65/90/150 (60)	.86	.57-.75	.80-.87
	1986	60/90/150 (60)	.86	.56-.75	.81-.87

Table A2 (continued). AP Reliability estimates by Item Types

Test	Year	Max Possible Score MC/CR/Comp. (%CR in Comp)	MC Reliability	CR Reliability	Composite Reliability
American History	1982	100/30/180 (50)	.89	.60-.79	.84-.90
	1983	100/30/180 (50)	.89	.52-.79	.83-.90
	1984	100/30/180 (50)	.90	.54-.79	.84-.90
	1985	100/30/180 (50)	.89	.54-.79	.83-.90
	1986	100/30/180 (50)	.90	.49-.79	.83-.90
European History	1982	90/30/135 (67)	.90	.46-.63	.71-.79
	1983	100/30/180 (50)	.89	.42-.63	.79-.85
	1984	100/30/180 (50)	.90	.46-.63	.80-.85
	1985	100/30/180 (50)	.91	.44-.63	.83-.87
	1986	99/30/180 (50)	.90	.48-.63	.81-.86
Biology	1982	120/45/150 (50)	.93	.60-.85	.87-.94
	1983	120/45/150 (50)	.93	.68-.85	.89-.94
	1984	120/45/150 (50)	.93	.70-.85	.90-.95
	1985	120/45/150 (50)	.93	.66-.85	.88-.94
	1986	120/45/150 (50)	.93	.73-.85	.89-.93
Chemistry	1982	79/88/160 (55)	.90	.77-.95	.90-.96
	1983	84/88/160 (55)	.91	≥.724	≥.882
	1984	85/88/160 (55)	.90	≥.756	≥.889
	1985	79/88/160 (55)	.92	≥.790	≥.909
	1986	80/88/160 (55)	.91	≥.776	≥.905

Table A2 (continued). AP Reliability estimates by Item Types

Test	Year	Max Possible Score MC/CR/Comp. (%CR in Comp)	MC Reliability	CR Reliability	Composite Reliability
Mathematics Calculus AB	1982	45/63/126 (50)	.86	.80-.87	.91-.93
	1983	45/45/90 (50)	.90	≥.793	≥.916
	1984	45/45/90 (50)	.89	≥.796	≥.914
	1985	45/50/108 (50)	.89	≥.843	≥.927
	1986	45/54/108 (50)	.90	≥.848	≥.931
Mathematics Calculus BC	1982	45/63/210 (50)	.85	.79-.85	.90-.92
	1983	45/45/126 (50)	.87	≥.735	≥.886
	1984	45/45/90 (50)	.87	≥.711	≥.890
	1985	45/54/90 (50)	.88	≥.753	≥.897
	1986	45/54/108 (50)	.88	≥.804	≥.909
Physics B	1982	70/105/210 (50)	.91	.79-.95	.92-.96
	1983	70/90/180 (50)	.88	.80-.98	.91-.96
	1984	70/90/180 (50)	.90	.85-.96	.93-.97
	1985	69/90/180 (50)	.89	.86-.98	.93-.97
	1986	70/90/180 (50)	.90	.84-.98	.93-.97
Physics C Mechanics	1982	35/45/90 (50)	.85	.80-.96	.90-.94
	1983	35/45/90 (50)	.85	.71-.97	.87-.94
	1984	35/45/90 (50)	.85	.79-.97	.89-.95
	1985	35/45/90 (50)	.84	.70-.97	.87-.95
	1986	35/45/90 (50)	.87	.70-.97	.88-.95

Table A2 (continued). AP Reliability estimates by Item Types

Test	Year	Max Possible Score MC/CR/Comp. (%CR in Comp)	MC Reliability	CR Reliability	Composite Reliability
Physics C Electricity & Magnetism	1982	35/45/90 (50)	.88	.80-.95	.91-.95
	1983	35/45/90 (50)	.85	.77-.98	.89-.95
	1984	35/45/90 (50)	.86	.67-.98	.86-.96
	1985	35/45/90 (50)	.83	.74-.98	.88-.95
	1986	35/45/90 (50)	.86	.74-.98	.89-.95

REFERENCES

- Allen, N. A., Kline, D. L., & Zelenak, C. A. (1996). *The NAEP 1994 technical report*. National Center for Education Statistics.
- Camara, W.J., Kimmel, E.K., and colleagues (1997). *High School Assessment Design: Interim Report to the Maryland State Board of Education*. New York: The College Board.
- Carlson, J. E. (1968). *Effects of differential weighting on the inter-reader reliability of essay grades*. Unpublished Dissertation, University of Alberta.
- College Entrance Examination Board (1988). *The college board technical manual for the advanced placement program*. New York College Entrance Examination Board.
- Gulliksen, H. (1987). *Theory of mental tests*. (Reprint of work originally published in 1950). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kolen, M. J. & Brennan, R. L. (1995). *Test equating methods and practices*. New York: Springer.
- Wang, M. W., & Stanley, J. C. (1970). Differential weighting: a review of methods and empirical studies. *Review of Educational Research*, 40, 663-705.



Designing End-of-Course High School Assessments

**A Final Report to the Maryland State
Department of Education**

Volume II



Wayne Camara
Howard Everson
Robert Majoros
The College Board

Kathleen O'Neill
John Fremer
James Braswell
James Carlson

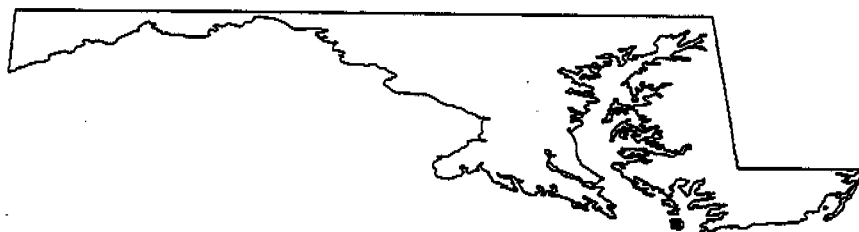
Walter Jimenez
Ernest Kimmel
Patricia Klag
Michael Lapp
Timothy Ligget
Karen Nulton
Marlene Supernavage
Janet Waanders
Ann Marie Zolandz
Educational Testing Service

August 1997

Designing End-of-Course High School Assessments

**A Final Report to the Maryland State
Department of Education**

Volume II



Wayne Camara
Howard Everson
Robert Majoros
The College Board

Kathleen O'Neill
John Fremer
James Braswell
James Carlson

Walter Jimenez
Ernest Kimmel
Patricia Klag
Michael Lapp
Timothy Ligget
Karen Nulton
Marlene Supernavage
Janet Waanders
Ann Marie Zolanz
Educational Testing Service

August 1997

CONTENTS

	Page
Appendix C 1. Working assumptions about the HSA tests.....	C-1
2. Test specifications for all content areas	C-5
Appendix D Illustrative test questions for all content areas	D-1
Appendix E 1. Issues and recommendations from the Program Specifications Committee	E-1
2. Issues and recommendations from the Test Administration Specifications Committee	E-7
3. Rosters of the specifications committees	E-15
Appendix F 1. Topics for a Maryland booklet on test security.....	F-1
2. Flowchart detailing one way of doing cheating analyses	F-3
3. Effects of scorer reliability: Examples from College Board testing programs.....	F-5
4. <i>Why and How Educational Testing Service Questions Test Scores</i>	F-11
5. <i>ETS Sensitivity Review Guidelines</i>	F-21
6. <i>ETS Standards for Quality and Fairness</i>	F-25
7. <i>Code of Fair Testing Practices in Education</i>	F-69
Appendix G Additional issues concerning the use of calculators on HSA	G-1

Appendix C

Working Assumptions About the HSA Tests

[ETS/CB comment: These are the original working assumptions presented to the various committees. Some of these working assumptions have since been changed and the modified assumptions have been incorporated into the specifications.]

Definition of working assumption: A set of assumptions that provide a common understanding so that we can begin the task. These assumptions are not final. Anyone who feels that an assumption is inappropriate should comment to MSDE or ETS staff. MSDE will then ensure that the issue is referred to the appropriate group for resolution.

Test Specifications Assumptions:

1. Test length: 3 hours, of which 15 minutes is allocated for instructions.
2. Test design:
 - Combination - Math, Science, Social Studies
 - Preparation Plus - English
3. Preparation Plus
 - a. 75 min. of activity covered within 3 days
 - b. Materials administered within 1 week of test
 - c. Students will prepare individually.
 - d. Activities will occur primarily in the classroom.
 - e. Classroom activity will be primarily teacher led.
 - f. There will be a high level of scripting so all teachers can emphasize the same information.
 - g. There will be a new activity for each test form (i.e., a new activity for the operational and the make-up forms).
4. Item type distribution:
 - 90 selected response items (taking 90 minutes)
 - 45 minutes of brief constructed response items
 - 30 minutes of extended constructed response items
5. Each test form will measure every CLG and every expectation, to the extent possible. Every indicator will not necessarily appear in each test form since indicators have been included as illustrations of content.

6. Students who are adequately prepared should be able to pass the test in the time allotted. Most examinees should be able to finish the test in the time allotted.
7. There will be approximately equal gender balance in each test; a minimum of 15% of the test material involving people (reading passages, other textual material, names, visual material) will portray minority topics.
8. The tests will be given in English, and students will respond in English.
9. All constructed response (CR) questions will be scored using a generic rubric with a full behavioral scale. A CR score can contribute only once to the total test score. Other decisions (such as the range of the score scale and whether the generic rubric needs to be supplemented) will be decided by a scoring team.
10. Some diagnostic information, called "goal scores," should be available for at most 2-3 areas to help in remediation.

Program Specifications Assumptions

Cross-test Issues:

1. All tests must be the same length.
2. In Math, Science, and Social Studies, test difficulty must be the same across all tests. In English, the test will increase in difficulty from test 1 to test 3.
3. Scratch paper must be provided for students during the tests. Students may not bring scratch paper into the testing room, nor take scratch paper from the testing room.
4. Testing terminology (e.g., explain, justify, analyze) must be defined for the students within the test.
5. Students may take the operational test as often as it is offered within the same year (i.e., in January and in June).

Student Populations:

6. Only students working toward a MD high school diploma are expected to take the HSA tests.
If LEP students are working toward a MD diploma, they will be required to take the HSA exams.

7. Accommodations for students with disabilities will be offered according to “Requirements for Accommodating, Excusing, and Exempting Students in Maryland Assessment Programs.”

Specifications writers should check that this document covers all possibilities which may arise regarding accommodations or exemptions. For unique situations which are not addressed in this document (e.g., Preparation Plus), it will be necessary for the specs writers to create new specifications.

8. Any modifications/accommodations that are specified in the student’s IEP must also be provided in the activities associated with the test (Preparation Plus, testing situation, testing environment).
9. All students must meet the same passing level regardless of disability or language proficiency.

Test Administration Assumptions

1. All materials (such as equipment) that are specified in the Core Learning Goals for instructional purposes are also subject to use for assessment purposes.
2. Within an individual test form, selected response questions may be intermingled in blocks with constructed response questions.
3. Scratch paper must be provided for students during the tests. Students may not bring scratch paper into the testing room, nor take scratch paper from the testing room.
4. All tests will be given during the normal school week (i.e., Monday through Friday).
5. All tests will require 180 minutes for administration: 165 minutes of actual testing and 15 minutes of instructions and directions.
6. All materials appropriate to the test (including, but not limited to, pens, highlighters, consumable texts) must be provided during the administration of the test.
7. For tests using Preparation Plus, only print materials will be available during test administration.

General Specifications for the Maryland HSA Tests

Test length will be three hours for each test. Of the 180 minutes, 15 minutes will be allocated for directions, leaving 165 minutes of testing time.

Reviews for multicultural and fairness concerns must be conducted at both the item and test level. At the item level, these reviews should ensure that the materials reflect Maryland's multicultural nature. The reviews should also reflect the Board of Education's commitment to develop test materials that are free of racist, sexist, or otherwise potentially offensive language and images. To this end, materials which involve stereotyping, inflammatory or highly controversial topics, or inappropriate tone should be considered unacceptable. At the test level, these reviews should also consider the balance and overall impression of the test.

There will be four answer choices for the selected-response items. Selected-response test items should not include "all of the above" or "none of the above" as answer choices. Selected-response test items should not include Roman numeral (or "k-type") items.

Selected-response items will be administered in blocks. That is, students will not be required to switch from one item to another between the selected-response answer sheet and the constructed-response answer booklet.

Test specifications will be developed to limit the number and location of visuals that are printed in half-tone or color.

Item writers must indicate how each item links to one or more Skills for Success. The item pool should contain items which link to all testable Skills for Success.

Sports scenarios will not be used as a context for items in HSA examinations.

Dictionaries will be available to students throughout the tests where they are permitted. Students should be aware that any time they spend using dictionaries will impact the amount of time they have to complete the test.

A small number of subscores may be developed for each test, depending on results from the no-fault administration.

Specifications for the Maryland HSAs in English

Introduction

This document specifies the testing approach, timing, organization, content and recommended item formats for Tests 1, 2, and 3 in English. The English content committee that assisted in preparing the specifications expects the tests to be aligned as closely as a standardized instrument can be with the English Core Learning Goals.

The overall approach to the tests in English is a model called Preparation Plus, which includes guided reading time outside of the timed test. A rationale for the model and a brief summary of the discussion about the challenges posed by the model follow.

Following the rationale and summary are General Specifications and Specifications for Tests 1, 2, and 3, accompanied by the portions of the Core Learning Goals, Expectations, and Indicators each test must cover.

Rationale for Preparation Plus

The introduction to the English Core Learning Goals states that "Reading, writing, speaking, and listening require the learner to engage in preparatory activities and then to construct meaning, compose, and evaluate." We believe that these "preparatory activities" must be incorporated into the Preparation Plus phase of the HSA English tests.

"In the English classroom students interpret, generate, and evaluate texts (Preface to English CLGs). In Preparation Plus we envision that students will be able to interact with texts (print or non-print) such that they read or view them, interpret them, and evaluate them. We acknowledge that students possess a wide range of experiences with print and non-print materials and that "their experiences and backgrounds influence their understanding of the text" (Introduction to English CLGs). Certain experiential variables, therefore, cannot be controlled in the Preparation Plus phase. To mitigate these variables as much as possible, the experience with text during the preparation phase should be contained and specifically scripted to provide students an opportunity to reflect not only on the text but also on the skills and processes they have mastered during their courses.

Reading a text in and of itself is not sufficient for understanding the richness of a text. Students need time to construct, examine, and extend meaning, including time for close reading, annotations, and reflective journals. Preparation Plus builds time and offers guided exercises so that students may engage in these enriching activities.

By providing preparation immediately prior to testing, this prep plus option forestalls out of class investigation. Prep Plus benefits all students, but most particularly LEP and Special Ed students by allowing more time for reflection and by shortening the daily block of time spent in test-taking.

Preparation Plus models the process of English instruction not only by providing opportunities for students to use before-, during-and after-reading strategies, but also by differentiating between surface reading and interpretation or analysis of a text.

Summary of Discussion about Administration of the Preparation Plus Model

The rationale above was prepared because the English Content Team wished to make clear the importance of the preparation period in the light of the challenges the model will pose for administration. The first model the team considered was one in which students would encounter a text or texts in their classrooms in the days prior to taking the English Test. Concerns about the model, raised first by the Content Group itself, and reinforced later by the Program Specifications and Test Administration Committees, included: security of the pre-released texts; unequal access to textual aids between the time of encountering the preparatory texts and the testing time; scheduling of the preparatory activities during regular classroom hours; provision of a comparable preparatory activity for instances of student or teacher absence; inequitable administration of the preparatory portion. Problematic though these issues are, the English Content Group remained unanimous in its desire to keep the preparation portion of the English tests. The committee therefore recommended a change to both the preparatory portion of the test and the test itself.

Many of the problems surrounding the original preparation period had to do with the lapse of time between when the students were initially to encounter the texts and when they were to be tested on them. To address these concerns while keeping true to the spirit of the preparation period as a time of quiet reading and reflection, the English Content Group suggests the following: 1) administration of the English test over two days; 2) a sixty minute preparatory period attached to the first day of testing. The revised English test, then, took the shape outlined on the first page of this document.

The two drawbacks that the English Content team foresees in this new iteration of prep plus are 1) increased absence vulnerability, and 2) the perception of an increased total testing time. While the team is aware that administering a test on two days rather than one poses administrative and logistical concerns, it feels that the benefits gained by allowing students--*who are being tested on their ability to read and analyze documents*--the time necessary to perform to the best of their abilities outweighs the administrative drawbacks. As far as the second issue is concerned, the team would emphasize that the length of the test has not increased; students are still only being tested for three hours. What has changed is that the preparation portion--a block of time to read and analyze texts -- will be encountered just prior to the first portion of the timed test.*

* The Content group recognizes that the test administration and program specification committees recommend that all tests be administered on one day. The group still recommends a two-day administration of the English tests to mitigate testing fatigue. Should this administration prove impossible, the group acknowledges that the English tests could be administered on one day. This one day administration assumes that the portion of the test labeled "Day 1" is administered in the morning and the portion of the test labeled "Day 2" is administered in the afternoon.

General Specifications for English Tests 1, 2, 3

I. General Policies

- A. The tests will be taken in **sequential order** (1,2,3) except where exemptions are granted for special circumstances (transferring into the state, for example).
- B. It is expected that the Preparation portion of the test will be **administered** by **certified English** teachers when possible.
- C. Teachers administering the preparation portion of the exam are expected to be familiar with the materials; these teachers should be permitted to **examine the preparation** portion of the exam at least one day prior to the test itself.
- D. **Dictionaries and thesauruses** will be available to students throughout the preparation and testing portions of the tests. MSPAP guidelines for selecting and providing these aids should apply. Students should be aware that any time they spend using dictionaries will impact the amount of time that they have to complete the tasks covered in the test.
- E. All Selected Response (SR) items assume a **testing rate** of one minute per item. All Selected response (SR) items will be four-option items in an A-D format. No more than 3-4% of items are to have negative wording in the question stimulus ("stem"), that is, wording such as "not" or "except." Sets of items tied to a single stimulus (or to a cluster of stimuli) will comprise approximately 75% of the test.
- F. Directions are expected to be succinct and clear. Directions for SR items are to be accompanied by a completed sample item. Directions and sample items will be repeated each time students are directed to begin a new item type. If one item type is found in more than one section of the test, the directions and sample item will be repeated.
- G. While numbers of questions are suggested for the indicator level, it should be recognized that this level is intended to reinforce the expectation level. The indicators can be seen as a general guide to the degree of importance that should be attached to individual indicators and not necessarily as definitive. Where, for instance, only one out of a number of indicators is assigned to a particular test and testing this indicator exclusively would violate the intended purpose of the goal and expectation covered, items referring to indicators covered in previous tests are acceptable. Test 2, then, could occasionally sample the indicators already covered in test 1, and test 3 could sample all of the indicators for a given expectation. It is not acceptable for tests to draw from indicators covered in future tests. The primary emphasis should always be on the assigned indicators; others should only be sampled when doing so helps to emphasize the core of the expectation and goal covered.

- H. MSPAP guidelines regarding what can and cannot be posted in classrooms where tests are being administered should be followed.
- I. Special conditions for special populations will follow MSPAP and/or other state guidelines. It should be noted, however, that, while oral stimuli should be transcribed for deaf and hard-of-hearing students, the questions or topics posed about those texts may include some small portion that are dependent upon hearing the text. Visual stimuli are for the most part not likely to be Brailleable (stimuli from film; works of art; cartoons, and the like). Where they can be brailled, they should be.

II. Equipment

The Core Learning Goals in English require the use of multimedia technologies. Because of this, the English Content Group—while recognizing that equitable access to sophisticated equipment may be problematic in the short term--includes alternative methods of textual interaction (including but not limited to film, tapes, projections, CD Roms) in the test specifications. That the specifications can be met without relying on these more sophisticated methods of textual presentation does not obviate the need for these alternative methods to be included in the test at the earliest date possible.

III. Definitions of Terms

A. TEXT

The word “text” is used throughout the specifications to refer to a broad range of documents, including: a) written (i.e. fiction and non-fiction prose, poetry, and drama); b) print (e.g. photos, film stills, print reproductions of artwork and cartoons); and c) non-print (i.e. visual texts such as film and oral texts such as taped speeches, literary readings, and recorded lyrics). When the word “text” is intended to convey a more limited sense—as, for example, when it is meant to refer specifically to a prose passage—this will be clearly delineated. In the absence of such delineation, the English specifications refer to the broader definition.

B. AUTHENTIC TEXT

The term “authentic text” is meant to describe a text composed for a general readership and not composed specifically for the test. When they are based on written texts, such “authentic texts” are expected to utilize a portion of text that is largely unchanged from its original incarnation. Shakespeare’s words would be based on an acceptable Folio edition, for example, and not modernized or transmuted in any way for testing purposes. Texts can and should be modified, however, in order to a) present students with a brief (one or two sentence) introduction to help situate them in the text, and b) define/annotate a limited number of words that might otherwise present unnecessary barriers to understanding. “Authentic” texts can be retyped to fit

testing pages. Given these modifications, it is expected that the texts encountered on the test should be accessible to most students. Where heavy annotating or footnoting becomes necessary to make a text accessible for most students, this text will be deemed inappropriate for testing purposes.

C. PREPARATION PLUS

The term “**preparation plus**” signifies the structure chosen for English Tests 1,2, and 3. These tests rely on a 60 minute reading and preparation period prior to testing. A more detailed rationale for and description of preparation plus can be found in the part of the document called “Rationale for Preparation Plus.”

D. PREPARATION PERIOD

The term “**preparation period**” (or “**preparation portion**”) is used to signify the 60 minute reading period which precedes the administration of the English tests.

E. SCRIPTING

The term “**scripting**” refers to an actual script that will be composed for and read by the teachers administering the preparation portion of the exam. This script will: 1) guide students through the reading process; 2) define when and how the teacher may read sections of the preparatory material to the class; 3) explain how teachers may suggest before-, during and after-reading strategies to their students. The script is expected to be followed verbatim by every teacher who oversees the preparatory portion of the test. The developer may propose as an alternative a taped script that paces the preparatory period and includes narrated instructions and readings of any stimulus material that is to be presented orally.

IV. General Stimulus and Item Distribution Guidelines

- A. The **texts selected** are intended to represent a wide variety of historical and cultural milieus as well as a variety of genres.* As a general guideline, we assume:
1. Texts drawn from literature should be of recognized literary merit.
 2. Texts selected should represent equally male and female authors.
 3. Translated texts are acceptable.
 4. A variety of genres will be represented across the three English tests, though the majority of stimuli will be drawn from prose genres.

* The Content group was resistant to narrowing the text specifications more specifically with regard to percentages of texts tied to genres, time periods, or origin of texts (i.e. US versus non-US authors, native versus translated English texts). The group recommends that texts should be accessible to the majority of students; they wish to rely on field-testing to determine if individual texts are accessible and appropriate.

5. Test assemblers should attempt to balance more challenging texts (possibly those prose or poetry pieces written in archaic language, or certain abstract art) with more accessible texts (possibly a modern work of prose or poetry).
 6. Texts and items explicitly representing minority experiences or written by minority authors must account for at least 15% of the texts on the test. In selecting these minority texts, efforts must be made to take into account the specific minority groups found in MD public schools. These minority groups should not, however, represent a definitive selection of acceptable minority representation—efforts should be made to embrace minority groups from outside MD as well.
- B. All SR items that test writing ability are expected to be **generally balanced** across the following subject areas: 1) humanities (music, philosophy, literature, art); 2) practical affairs (school and work related topics, day-to-day activities); 3) human relations (interpersonal relationships, emotions); and 4) other academic content areas.
- C. The test developer will present a system of covering **testable errors** designed to test aspects of language skills. One possible list would include the following errors: subject/verb agreement; tense; connectives; modifiers; pronoun usage; diction; idioms; parallelism; sentence fragments; wordiness; punctuation, capitalization; and sentences that contain no error. Once the master list is accepted by MSDE, the test developer will work to ensure that there is general balance of error types in questions of language use.

V. Specific Description of Item Types

A. SENTENCE CORRECTION

This item type is intended to present students with sentence level errors. Items should offer students alternatives at the phrase rather than the word level. One way of accomplishing this is to present students with sentences that are partially or wholly underlined. Students can be asked to choose the best version of the sentence from among a number of possible revisions. This section is primarily intended to test students' abilities to "demonstrate the ability to control language by applying the conventions of standard English in writing and speaking" (CLG 3). The item type(s) should present students with plausible errors; students should be required to revise sentences using integrated rather than isolated language skills. For Test 1, each item stimulus is expected to contain a maximum of 20 words; for Test 2, each item stimulus is expected to contain a maximum of 23 words; for Test 3, each item stimulus is expected to contain a maximum of 26 words.

B. REVISION IN CONTEXT

This item type is intended to test student's "ability to control language by applying the conventions of standard English" (CLG 3) within the context of a paragraph or

group of paragraphs. While students may be asked to identify sentence level errors as part of this item type, the item is primarily envisioned as a place in which students use the revision skills that they would normally use when editing a draft of a paragraph or essay. Questions should focus on the ways in which the “structure of language, including grammar concepts and skills” are used to “strengthen control” of written texts that are longer than a sentence (CLG 3). One way of accomplishing this is to present students with an essay (described as a draft of a student paper) which contains a number of errors. Students could be asked to revise the paragraph for clarity and organization. While specific items within this item type may ask students to revise a single sentence for clarity, what must be stressed is the importance of this clarity to the paragraph/essay as a whole. The emphasis of this item is intended to be on larger than sentence level issues of coherence, structure, and revision choices. For Test 1, each item stimulus is expected to contain a maximum of 20 words; for Test 2, each item stimulus is expected to contain a maximum of 23 words; for Test 3, each item stimulus is expected to contain a maximum of 26 words.

C. CONSTRUCTION SHIFT

This item type is intended to test a student’s ability to revise sentences so that they are appropriate for new audiences or purposes. Students CAN be asked to select “language appropriate for a particular audience and purpose” (CLG 2). The ability of students to manipulate language, rather than to identify specific misuses of language, is what is paramount in this item type. One way of testing this skill is to present students with complete correct sentences. Students can then be presented with a one or two word stimulus, and asked to rewrite the sentence including the stimulus in their revisions. For Test 1, each item stimulus is expected to contain a maximum of 20 words; for Test 2, each item stimulus is expected to contain a maximum of 23 words; for Test 3, each item is expected to contain a maximum of 26 words.

D. MAKING MEANING FROM A SINGLE STIMULUS

This item type is intended to test students’ ability to make meaning from single texts. Authentic texts, then, should serve as the primary stimulus for the items which follow them. Students can be asked to “demonstrate the ability to respond to a text by employing personal experiences and critical analysis” (CLG 1). This item type is not intended to test students’ control of the conventions of English, but is meant to focus students’ attentions on their ability to “construct, examine, and extend meaning of traditional and contemporary works” (CLG 1, 2). One way of accomplishing this is to ask students questions based on their readings of a single text. All items should strive to reinforce the process of analysis and reflection generally expected from students as part of their strategies for implementing “before, during, and after reading, viewing, and listening” skills (CLG 1, exp.1). This item type is meant to elicit higher-order cognitive skills such as reasoning and analyzing, and is not intended to reward the less challenging ability simply to search for answers already contained in the passage(s). Where the meanings of individual words in passages are asked for, the

correct answers should rely on students' abilities to contextualize the words. Testing a student's familiarity with single words isolated from the passage is inappropriate; testing a student's ability to extrapolate the meaning of an unusual word or phrase *as it is used in the context of the passage* is appropriate.

E. MAKING MEANING FROM PAIRED STIMULI

All of the expectations for "MAKING MEANING FROM A SINGLE STIMULUS" apply to this item type; rather than testing students ability to extrapolate meaning from a single text, however, this item type is intended to test students' abilities to make meaning from two texts as they relate to each other. One way of accomplishing this is to ask students to consider and answer questions about two short related texts. Since the goal of this item type is to test how well students can synthesize information from two texts that are somehow related, at least 35% of the questions should address the works in relationship to each other. Appropriate questions might direct students to: 1) examine x in passage 1 as it relates to y in passage 2; 1) describe the basic relationship between passages 1 and 2 (complimentary, antagonistic, etc.); 3) examine a phrase that is used differently in passage x than it is in passage y. As with the previous item type, this item type is meant to elicit higher-order cognitive skills such as reasoning and analyzing, and is not intended to reward the less challenging ability simply to search for answers already contained in the passage(s).

F. PREPARATION-BASED BRIEF CONSTRUCTED-RESPONSE ITEMS

This item type is intended to test students' abilities to analyze the materials presented to them in the preparation portion of the assessment. These BCRs must flow directly from the texts and preparatory activities, and should attempt to elicit evidence of students' abilities to make meaning from the texts. While this item type can assume that the student has had time to interact with the texts found in the preparation portion of the exam, not questions can refer explicitly to any writing the students may have done in the preparatory portion. That is, while students may be encouraged to use reading strategies that may include written responses during the preparatory period, no questions can ask students to return to any specific writing that they may have done. It is acceptable to assume that the extra time that students have spent analyzing the texts can lead to a deeper level of engagement with the texts such that questions may be more complex than they would had the students only a few minutes to read the stimulus material. It is not acceptable to assume that any of the notes that students may have committed to paper in the preparatory period stand as the sole basis of a BCR item. Each BCR will elicit a slightly different response based on the CLG, expectation, and indicator that it addresses. The breakdown of these goals, expectations, and indicators and their effects on individual BCRs will be indicated at the test level. Formats for the BCR response may vary. Some possibilities include a prose response; a chart; a sentence or two of introduction followed by a list; a sketch (chiefly, though not exclusively, intended to elucidate questions of dramatic

intent/staging/ lighting and other related topics); poetry. There are three lengths of BCRs called for: 1) five minute; 2) ten minute; 3) 15 minute. In general, where a prose response is required students will be expected to respond to five minute BCRs in two or three sentences; to 10 minute BCRs in a paragraph; and to 15 minute BCRs in an extended paragraph or several paragraphs. Where responses other than prose are required, their length and complexity should require a level of sophistication comparable to those guidelines just described.

G. NON-PREPARATION-BASED BRIEF CONSTRUCTED-RESPONSE ITEMS

This item type is intended to test students' abilities to compose based on a stimulus of one or more than one short text(s). The length of the stimuli will vary by test, and will be indicated at the test level. Each BCR will elicit a slightly different response based on the CLG, expectation, and indicator that it addresses. The breakdown of these goals, expectations, and indicators and their effects on individual BCRs will be indicated at the test level. Formats for the BCR response may vary. Some possibilities include a prose response; a chart; a sentence or two of introduction followed by a list; a sketch (chiefly, though not exclusively, intended to elucidate questions of dramatic intent/staging/ lighting and other related topics); poetry. Students are expected to answer this item in five minutes; in general, where a prose response is required students will be expected to respond in two or three sentences. Where responses other than prose are required, their length and complexity should require a level of sophistication comparable to a prose answer of this length.

H. EXTENDED-CONSTRUCTED RESPONSE ITEMS

This item type is intended to test students' abilities to "compose in a variety of modes by developing content, employing specific forms, and selecting language appropriate for a particular audience and purpose." (CLG 2) While the primary emphasis of this item type is on the student's ability to compose in prose, CLG 2 specifically requires that students be able to "compose in a variety of modes." To embrace the "variety of modes" that a thoughtful response to a prompt may take, the Tests 1 and 3 leave the choice to respond in non-essay format (i.e. poetry, drama, verse) open to the individual student. In test 2, which covers CLG 2, exp.1, ind. 2 ("The student will compose to inform by using appropriate types of prose"), only the essay format is acceptable. IN ANY MODE, THE STUDENT'S RESPONSE IS EXPECTED TO BE A DRAFT. The directions should clearly direct students to compose a draft, and the scoring rubric must clearly state that the responses are expected to be drafts. The essay prompt is expected, in general, to be no more than a few sentences. It should direct students to their own personal, academic, and/or literary experiences as the primary basis for their responses. Outside knowledge of literary, historical, or cultural events--while it may enhance individual responses--should not privilege certain responses over others. The topic must be broad enough that the majority of students who are prepared to compose a response to the topic can do so; virtually all

students should have something to say about the topic so that a sample of their writing may be obtained.

VI. General Description of Organization and Scripting of Preparation Period

- A. For all tests, the preparation period will be scripted in order to standardize the experience across schools and districts. This scripting may vary from test to test depending on the nature of the preparatory texts. If, for instance, a film is part of the preparation period, scripting may require teachers to pause on a certain shot and ask students to take note of camera angles, use of lighting or props, spatial relationships of certain characters, etc. If a poem is used, teachers may be instructed to read the poem aloud a set number of times to enhance student interaction with the text. In every event, scripting is designed to help students reflect on their individual interactions with text while mitigating any inequities that could arise from more impromptu student/teacher dialogues. The interaction sought is one between individual students and text.
- B. The scripted portion of the preparation period is intended to stimulate prior knowledge as a before-reading strategy. 20-30 minutes of the 60 minute reading time should be spent in non-scripted reading and analysis of the texts. This time is intended to allow students to read and focus at their own paces and in their own ways on the texts presented to them. A minimum of 2 and a maximum of 4 texts should be presented to students. These texts should be of a length that the students can comfortably read in 20-30 minutes. The intent is to allow students to interact with a piece such as a short story or lengthier article/poem/film that they would not have time to focus on in the timed portion of the test.
- C. Passages will be presented in the order their content and the test questions suggest. A passage providing historical or cultural context for a poem, for example, might precede or follow the poem, depending on whether students are expected first to make sense of the poem as an isolated piece or to consider the poem as a product of time, place, and author's experience.
- D. Scripting should reflect good instructional practices and call upon before-reading, during-reading, and after-reading strategies. Scripting should provide opportunities for interaction between student and text with advanced organizers such as knowledge charts. As part of the Scripting should encourage students to annotate and/or highlight the text as part of the strategic reading behaviors. Questions, small tasks (such as filling in a chart), or guidance on note-taking should lead students through the texts while allowing room for differences in styles of thinking and working. An organizational chart, for instance should be accompanied by some blank space and the invitation to make notes in a format of the student's choosing. Questions meant to help students discover meaning and the textual evidence for that meaning should not force a single narrow interpretation. The emphasis in the preparation period is on the student's interaction with the texts(s). All scripted suggestions/activities are intended

to guide the student to his or her own interpretation of the text(s), not to a specific, scripted interpretation. The material used should be available to students during the timed portion of the test.

ENGLISH TEST 1

English Test 1	% Testing Time* (165 min)	Link to Skills for Success
Goal 1	28	
E1		1.1-3, 2.2-2.4, 3.1-3, 4.1-2
E2		1.1-3, 2.1-2.4, 3.1-3
E3		1.1-3, 2.1-2.4, 3.1-3, 4.3
Goal 2	35	
E1		1.2, 2.1-2.2, 3.1-3.2
E2		1.2, 2.1-2.2, 2.4, 3.1, 3.3
E3		1.3, 2.2, 3.2
Goal 3	20	
E1		1.2, 1.3, 2.2., 3.1, 3.2
E2		1.2
E3		1.2, 2.2
Goal 4	17	
E1		2.2
E2		
E3		3.1, 3.3
TOTAL	100%	

* The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

TEST SUBSCORES:

Recommended subscores are:

1. Goal 1
2. Goal 2
3. Goal 3
4. Goal 4

DISTRIBUTION OF ITEM TYPES

Seven brief constructed response questions for a total of approximately 45 minutes.

One extended constructed response question of approximately 30 minutes.

Ninety selected-response questions for a total of approximately 90 minutes.

ENGLISH TEST 2

English Test 2	% of Testing Time* (165 minutes)	Link to Skills for Success
Goal 1	32	
E1		1.1-3, 2.2-2.4, 3.1-3, 4.1-2
E2		1.1-3, 2.1-2.4, 3.1-3
E3		1.1-3, 2.1-2.4, 3.1-3, 4.3
Goal 2	31	
E1		1.2, 2.1-2.2, 3.1-3.2
E2		1.2, 2.1-2.2, 2.4 3.1, 3.3
E3		1.3, 2.2, 3.2
Goal 3	19	
E1		1.2, 1.5, 2.2, 3.1, 3.2
E2		1.2
E3		1.2, 2.2
Goal 4	15	
E1		2.2
E2		
E3		3.1, 3.3
TOTAL	100%	

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

TEST SUBSCORES:

Recommended subscores are:

1. Goal 1
2. Goal 2
3. Goal 3
4. Goal 4

DISTRIBUTION OF ITEM TYPES

Seven brief constructed-response questions for a total of approximately 45 minutes.

One extended constructed-response question of approximately 30 minutes.

Ninety selected-response questions for a total of approximately 90 minutes.

ENGLISH TEST 3

English Test 3	% of Testing Time* (165 min)	Link to Skills for Success
Goal 1	31	
E1		1.1-3, 2.2-2.4, 3.1-3, 4.1-2
E2		1.1-1.5, 2.1-2.4, 3.1-3.4
E3		1.1-3, 2.1-2.4, 3.1-3, 4.3
Goal 2	33	
E1		1.2, 2.1-2.2, 3.1-3.2
E2		1.2, 2.1-2.2, 2.4, 3.1, 3.3
E3		1.3, 2.2, 3.2
Goal 3	16	
E1		1.2, 1.5, 2.2, 3.1, 3.2
E2		1.2
E3		1.2, 2.2
Goal 4	20	
E1		2.2
E2		
E3		3.1, 3.3
TOTAL	100%	

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

TEST SUBSCORES:

Recommended subscores are:

1. Goal 1
2. Goal 2
3. Goal 3
4. Goal 4

DISTRIBUTION OF ITEM TYPES

Six brief constructed-response questions for a total of approximately 45 minutes.

One extended constructed-response question of approximately 30 minutes.

Ninety selected-response questions for a total of approximately 90 minutes.

Specifications for the Maryland HSAs in Mathematics

SPECIFICATIONS:

Test items should be appropriate for and appeal to the age group of the students. Language in the items should include correct mathematical terminology, but should not restrict or obscure meaning.

In general, items should not require the use of manipulatives, except possibly for simulations. Generic manipulatives should be available to students during testing if they have used them for instruction and choose to use them for testing. Acceptable materials include algebra tiles, patty paper, miras, unlabeled 3-dimensional models. In addition, the following should be available to all students throughout the test: formula page, ruler, scratch paper, graph paper, and graphing calculator. The following materials are prohibited: labeled/named geometric figures, "how-to" posters that include steps for solving specific mathematics problems, constructions, etc. (These lists should be revisited as needed.)

Use of current technology is expected throughout the test. The mathematics tests will require a graphing calculator*. Students should have access to the graphing calculator that they regularly use during instruction. Therefore, no constraints or limitations should be placed on the type of graphing calculator used on the test. Questions should be written so that students who have graphing calculators that have more capabilities do not have an unfair advantage over those who have more limited calculators. At a minimum, calculators must have the capability to do the following:

- table functions
- point plotting
- linear fit
- solutions to systems of equations
- statistics: mean, median, mode, interquartile range
- maxima and minima
- trigonometric functions values
- matrices

*See individual tests for additional information.

It is recommended that MSDE provide generic directions for specific graphing calculator models. In addition, the Content Team recommends that students be given directions for rounding to the appropriate place value.

MULTICULTURAL CONCERNS:

Test materials should reflect multicultural representation that is consistent with the public school population of Maryland.

There should be gender balance in materials and references that contain people.

Items should be bias free with respect to multicultural representation, gender balance, and background knowledge. (Refer to MSDE's Bias and Sensitivity Document.)

ITEM SPECIFICATIONS:

Negative items: A maximum of 5% of the items on the tests should be negatively worded.

Item sets: No more than 50% of all items in the tests should be in sets.

Visuals: The number of items that should contain visuals and the nature of these visuals should be decided jointly by the content group and specialists in the field of teaching/testing visually impaired students.

PHASE-IN ISSUES: Technology acquisition and use

MATHEMATICS TEST 1

	% of Testing Time*	Link to Skills for Success
Goal 1		2.2.2.2, 2.2.2.3, 2.2.2.5, 2.2.2.6, 2.2.3.1-2.2.3.5, 2.4.3, 2.4.4, 3.1.3.3, 3.1.3.4, 3.1.3.6, 3.1.4.12, 4.2.3.3, 4.3.3.4, 4.4.2.3, 4.4.2.4
E 1.1	26.7%	
I1		2.1.1.7, 2.2.1.3-2.2.1.6
I2		3.2.5.1, 3.2.5.2
I3		
I4		2.4.1.1-2.4.1.4, 2.4.2.2-2.4.2.6
E 1.2	32.7%	
I1		2.2.1.3-2.2.1.6
I2		
I3		
I4		2.4.1.1-2.4.1.4, 2.4.2.2-2.4.2.6
I5		2.4.1.1-2.4.1.4, 2.4.2.2-2.4.2.6
Goal 3		2.2.1.3-2.2.1.6, 2.2.2.2, 2.2.2.3, 2.2.2.5, 2.2.2.6, 2.2.3.1-2.2.3.5, 2.4.1.1-2.4.1.4, 2.4.2.2-2.4.2.6, 2.4.3, 2.4.4, 3.1.3.3, 3.1.3.4, 3.1.3.6, 3.1.4.12, 3.2.4.2-3.2.4.4, 3.2.5.1, 3.2.5.2, 4.2.3.3, 4.3.3.4, 4.4.2.3, 4.4.2.4
E 3.1	21.2%	3.3.3.4, 3.3.3.5
I1		2.1.1.3, 2.1.1.5, 2.1.1.7, 2.1.2.3, 2.2.4.1-2.2.4.3, 2.2.4.6
I2		
I3		2.1.1.7
E 3.2	19.4%	
I1		2.1.1.6-2.1.1.8
I2		
I3		2.1.1.3, 2.1.1.4, 2.1.1.6-2.1.1.8, 2.1.2.3, 2.2.4.1-2.2.4.3, 2.2.4.6, 2.2.4.8, 2.2.5, 2.2.6
TOTAL	100%	

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

GENERAL NOTES:

A set is a single stimulus followed by a series of independent items. A single stimulus should be used for: 1) several BCRs and/or ECRs and 2) several selected response and SPR questions.

Listed below are families of indicators which have natural links to create sets of items:

1.1.1 → 1.1.2

3.2.2 → 1.1.2, 1.1.4, 1.2.1, 1.2.3, 1.2.4, 1.2.5, 3.1.1

3.2.3 → 1.2 (except 1.2.2)

3.1.1 → 3.1.2, 3.1.3

3.2.1 → 3.1.1, 3.1.3, 3.2.3

Calculator Use:

It is recommended that 10 percent of the questions on the test require a graphing calculator. (Note that not all of these questions will require the graphing features of the graphing calculator; some may require the scientific features only.) Up to 40% of the questions on the test should be questions for which the calculator may be useful or necessary. (This includes the 10% for which the calculator is required.) This recommendation should be revisited on a regular basis.

EQUIPMENT AND MATERIALS:

In addition to the materials allowable in the testing room, all students will need the following: ruler (metric and customary) appropriate to tasks, graphing calculator (it is assumed that the students know how to use it), random number generator (spinner, number cubes, random number table), scratch paper, graph paper, tear-out page with appropriate formulas.

DISTRIBUTION OF ITEM TYPES:

Extended Constructed Response (ECR)

The test will contain 4 ECRs for a total of approximately 40 minutes. Each ECR will measure one expectation but multiple indicators and may have multiple parts. Student responses should include an explanation/narrative portion in each ECR.

Brief Constructed Response (BCR)

The test will contain 11 BCRs for a total of approximately 30-35 minutes. The BCRs should not solely require a numerical response. At least one of the BCRs for each expectation should require a narrative response.

Student Produced Response (SPR)

The test will contain 10 SPRs for a total of approximately 20 minutes. It will be important to watch for “appropriate precision” questions in which the answer depends on problem context. The gridded answers should include a variety of types of answers (not too many fractions, decimal, integer etc.) and the full range of the grid should be used as much as possible.

Selected Response (SR)

The test will contain 60 SRs for a total of approximately 70-75 minutes. The content committee recognizes that number of selected response questions indicated for this test is inconsistent with the original working assumptions and current recommended technical specifications. The number of selected-response questions may need to be increased to 90 or more as a result of pilot and no-fault tests in order to meet psychometric requirements. In addition, more constrained, less open-ended constructed response items and scoring tools may be advisable in order to ensure high rater agreement rates.

ITEM SPECIFICATIONS:

Real world settings: 50% - 60% of the items on the test should be set in a real world context. (See General Specifications for item specifications for both mathematics tests.)

TEST SUBSCORES:

Subscores should be reported for:

- Goal 1, Expectation 1.1
- Goal 1, Expectation 1.2
- Goal 3, Expectation 3.1
- Goal 3, Expectation 3.2

MATHEMATICS TEST 2

	% of Testing Time*	Link to Skills for Success
Goal 2	29.8%	2.2.2.2, 2.2.2.3, 2.2.2.5, 2.2.2.6, 2.2.3.1-2.2.3.5, 2.4.3, 2.4.4, 3.1.3.3, 3.1.3.4, 3.1.3.6, 3.1.4.12, 4.2.3.3, 4.3.3.4, 4.4.2.3, 4.4.2.4
E 1		2.2.1.3-2.2.1.6
I 1		
I 2		
I 3		
I 4		
E 2	38.1%	2.2.1.3-2.2.1.6, 2.4.1.1-2.4.1.4, 2.4.2.2-2.4.2.6
I 1		
I 2		
I 3		2.1.1.7
E 3	32.1%	2.1.1.7, 2.2.1.3-2.2.1.6, 2.4.1.1-2.4.1.4, 2.4.2.2-2.4.2.6
I 1		
I 2		
TOTAL	100%	

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

GENERAL NOTES:

Each indicator must be addressed in a BCR, ECR, or both.

There is concern about the level of precision of instruments required in a question. Questions that require measurements/drawings/constructions should specify the degree of precision required. The tolerance level of a response would not be expected to have precision beyond 1/4" or 0.5 cm., but the tools used may have a greater degree of precision.

Calculator Use:

Appropriate uses of technology are expected throughout the test.

It is recommended that 10 percent of the questions on the test require a graphing calculator. (Note that not all of these questions will require the graphing features of the graphing calculator; some may require the scientific features only.) Up to 40% of the questions on the test should be questions for which the calculator may be useful or necessary. (This includes the 10% for which the calculator is required.) This recommendation should be revisited on a regular basis.

(Note that the above percentages are targets for Test 2. Currently, graphing calculators do not have as many applications in geometry as in algebra, but this is expected to change as technology becomes more advanced.)

EQUIPMENT AND MATERIALS:

In addition to the materials allowable in the testing room, all students will need the following: protractor (any type), ruler (metric and customary), compass (any type), graphing calculator (it is assumed students know how to use it), scratch paper, graph paper, tear-out page with appropriate formulas.

DISTRIBUTION OF ITEM TYPES:

Extended Constructed Response (ECR)

The test will contain 3 ECRs for a total of approximately 30 minutes. Each ECR will measure one expectation but multiple indicators and may have multiple parts with multiple levels of difficulty (that start with easier prompts and move to more difficult prompts). Student responses should include an explanation/narrative portion in each ECR.

Brief Constructed Response (BCR)

The test will contain 9 BCRs for a total of approximately 45 minutes. The BCRs should not solely require a numerical response. At least one of the BCRs for each expectation should require a narrative response.

Student Produced Response (SPR)

The test will contain 10 SPRs for a total of approximately 18-20 minutes. It will be important to watch for “appropriate precision” questions in which the answer depends on problem context. The gridded answers should include a variety of types of answers (not too many fractions, decimal, integer etc.) and the full range of the grid should be used as much as possible.

Selected Response (SR)

The test will contain 60 SRs for a total of approximately 70-75 minutes. The content committee recognizes that number of selected response questions indicated for this test is inconsistent with the original working assumptions and current recommended technical specifications. The number of selected-response questions may need to be increased to 90 or more as a result of pilot and no-fault tests in order to meet psychometric requirements. In addition, more constrained, less open-ended constructed response items and scoring tools may be advisable in order to ensure high rater agreement rates.

ITEM SPECIFICATIONS:

Real world settings: 40% - 50% of the items on the test should be set in a real world context.

TEST SUBSCORES:

Subscores should be reported for:

Goal 2, Expectation 2.1

Goal 2, Expectation 2.2

Goal 2, Expectation 2.3

Specifications for the Maryland HSAs in Science

LABORATORY PRE-REQUISITE FOR SCIENCE TESTS

The science test committees believe that it is important for students to have hands-on experience performing laboratory experiments in class and, therefore, laboratory work is a presumed pre-requisite to the HSA science tests. This laboratory component is based on the ideas that science is a hands-on activity and that these activities will help students make connections between the sciences and their daily lives. If students are to make science an important experience in their lives, they must actively participate in the process of science. In order to be scientifically literate, students must learn to: formulate hypotheses, design experiments, collect and interpret data, effectively communicate the information they have gathered, and argue their conclusions persuasively. Students can learn these skills if they are able to participate in a variety of laboratory exercises.

Laboratory work will also help students understand the concepts embodied in the Core Learning Goals, enrich student learning, and bring Maryland schools to the forefront of science education. It will raise the standards of teacher training, and it will modernize science classrooms by providing a baseline of equipment and reagents.

At the core of the laboratory component is a series of concepts that are essential to each of the science disciplines. These concepts might, for example, be developed from the indicators in each Core Learning Goal. The concepts would be a series of statements that not only help define what the laboratory work will accomplish, but the statements might also act as ways to evaluate what the students have accomplished. (For example, teachers might include in a lesson plan: "After completing this laboratory, students will be able to describe how the rate of an enzyme-mediated reaction is related to environmental temperature.")

The laboratory exercises are intended to teach concepts, but these concepts can be reached through a variety of activities. Although suggested activities are included in the Core Learning Goals and in some of the individual test specifications, these activities are not the only way to accomplish the goals.

The HSA Science tests will not evaluate the procedures used in the laboratory experiments; they will evaluate the student's knowledge of laboratory concepts. For example, students might be asked to draw conclusions based on data from a laboratory experiment, or they might be asked to explain a concept using examples from their laboratory experience. It would also be reasonable to ask students to design an experimental procedure to test a particular concept. Students would not be asked to recall results from memory or to describe a specific procedure.

DEFINITIONS OF SELECTED RESPONSE ITEM TYPES IN SCIENCE TESTS

Laboratory-set items are sets of items which have common stimulus material that refers to a laboratory situation. The items may assess laboratory skills.

Classification-set items are sets of items which refer to the same set of answer choices. They may also have some common stimulus material.

Technical passage sets are sets of items which involve interpreting and analyzing a technical passage, either from an actual publication or a passage written for the test.

Discrete items may be either single items or items in a set with a common stimulus (and which do not fall into the above categories).

Earth and Space Science

	Percent of Testing Time*
Goal 2	
E1	14
E2	10
E3	8
E4	20
E5	20
E6	9
E7	6
E8	13
TOTAL	100

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

Earth and Space Science Skills and Processes

Earth & Space Science	% of Testing Time*
Goal 1	
E1	15
E2	25
E3	not testable
E4	25%
E5	15%
E6	10%
E7	10%
TOTAL	100 %

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

Earth and Space Science
Links Between Goals 1 and 2

Goal 2	E1 Openness & Skepticism	E2 Experimental Approach	E3 Select Instruments	E4 Data Analysis	E5 Communi- cations	E6 Math	E7 Interdis- ciplinary
E1 I ₁		X	X	X	X	X	X
I ₂		X	X	X	X	X	X
E2 I ₁				X		X	
I ₂		X		X	X		
E3 I ₁		X		X	X		
I ₂			X	X	X		
I ₃	X	X	X	X	X		
E4 I ₁							
I ₂				X	X	X	
I ₃				X	X		
I ₄				X	X		
I ₅	X		X	X	X		
E5 I ₁	X			X	X		X
I ₂				X	X		
E6 I ₁	X			X	X		X
I ₂	X		X	X			X
E7 I ₁		X	X		X	X	
I ₂					X	X	
I ₃		X	X		X	X	
E8	X	X	X	X	X	X	X

Earth and Space Science Link to Skills for Success

The links between the Earth and Space Science Core Learning goals and the Skills for Success are made through Goal 1 - Skills and Processes. Shaded blocks indicate a match.

SCIENCE EXPECTATIONS	SKILLS FOR SUCCESS EXPECTATIONS																
	1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																	
1.2																	
1.3																	
1.4																	
1.5																	
1.6																	
1.7																	

DISTRIBUTION OF ITEM TYPES:

Brief Constructed Response (BCR)

Six BCRs for a total of approximately 30 minutes.

Extended Constructed Response (ECR)

Three ECRs for a total of approximately 45 minutes.

Selected Response (SR)

Ninety SR questions for a total of approximately 90 minutes. The SR items will include discrete questions, classification-set questions, laboratory-set questions or technical-passage set questions. Directions for each type of SR will be provided.

RECOMMENDATIONS ABOUT ITEMS AND MATERIALS IN TESTING ROOM:

The Committee recommends the following:

1. 45% of the SR and 100% of the BCR's and ECR's have a visual - (e.g. - data tables, data graph, data table, data chart, instrumentation [thermometer, rain gauge, seismic readings], maps, geologic formations, star patterns, and some color weather maps
2. 50% of the items be set based
3. no more than 5% of the items contain negative stems
4. the following materials be available during the testing: graph paper, metric ruler, protractor, compass, calculator (e.g., TI20 non-programmable), color pencils (4 - red, green, orange, blue; pencil sharpener)
5. Dictionaries of any type will not be permitted.

MULTICULTURAL CONCERNS:

There should be minority topics or references in 15% of test materials that contain people.
There should be gender balance in materials and references that contain people.

EQUIPMENT NEEDED FOR TEST:

TV, VCR, ruler.

TEST SUBSCORES:

Appropriate subscores are:

1. Expectations 1 and 2
2. Expectations 3 and 4
3. Expectations 5, 6, and 7

EQUIPMENT FOR CLASSROOM LABORATORY WORK:

adding-machine tape	glycerin	psychrometer
anemometer	graduated cylinders	radiation kit
aneroid barometer	(10, 50, 100, 1000 ml)	rain gauge
beads (black, red, yellow)	hot plate	ring stand (with ring)
beakers (250, 400 or 600 ml)	hydrochloric acid	ruler (metric)
bucket	hydrometer	satellite imagery access
Bunsen burner	light bulbs	scissors
c-clamps	lodestone	slinky
candles	magnesium chloride	sodium bicarbonate
conduction kit	magnesium sulfate	sodium chloride
coriolis kit	magnetic compass	spectroscope
density kit	magnets	stream table
drafting compass	magnifying lens	strontium chloride
Erlenmeyer flask (250 ml)	maps (U.S., world)	test tubes
file	matches	test tube rack
food coloring	medicine dropper	thermometer (-10° to 110°)
fossil brachiopod kit	meterstick	timer (clock, watch, etc.)
fossil kit	nichrome inoculating loop	triple beam balance
glass tubing	plastic bags	weather station
globes (celestial, relief)	plastic box (clear, with lid)	

TOPICS AND CONCEPTS FOR SUGGESTED CLASSROOM LABORATORY EXERCISES:

Spectral analysis -- Identifying unknown elements based on their spectra; analyzing stellar chemistry and relative age

Orbital mechanics -- Describing and applying gravitational forces to determine the relationship between the orbital radius and period of revolution of a planet (Kepler's laws)

Convection, conduction, and radiation -- Describing heat transfer in systems, associated with meteorological data

Greenhouse effect, global warming, climate data, and ocean currents

Plate tectonics, earthquakes and volcanism, epicenter location

Angle of insolation -- Explaining world climate and seasons

Tides, lunar phases, eclipses -- Explaining these via the relative positions of Earth, Moon, and Sun

Age of the Earth -- Exploring the interrelationship between political climate and scientific theory over time

Models and order of magnitude - Construction of a model that involves a change in order of magnitude (e.g. geological time scale, distances in the Solar System, Richter scale, humanity's place in the time continuum)

BIOLOGY

	Percent of Testing Time*
Goal 3	
E 1	16-19
E 2	19-22
E3	16-19
E4	12-15
E5	19-22
E6	9-12
TOTAL	100

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

Biology Skills and Processes

Test Name	Percent of Testing Time*
Goal 1	
E1	5
E2	25
E3	0
E4	35
E5	17
E6	6
E7	12
TOTAL	100

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

Biology Links between Goal 1 and Goal 3

Goal 3	E 1 Openness & Skepticism	E2 Experimental Approach	E3 Select Instruments	E4 Data Analysis	E5 Communi- cations	E6 Math	E7 Interdis- ciplinary
E1 1		X	X		X		X
2		X	X	X	X	X	X
3		X	X	X	X	X	X
E2 1		X	X		X		
2	X	X	X	X	X	X	X
E3 1	X	X	X	X	X	X	
2	X	X	X	X	X	X	
3	X	X	X		X		
4	X	X	X		X		X
E4 1	X	X			X		
2	X	X	X	X	X		
E5 1		X	X	X	X	X	X
2	X	X			X		X
3	X	X	X	X	X	X	X
4	X	X			X		X
E6	X	X	X	X	X	X	X

Biology Link to Skills for Success

The links between the Biology Core Learning goals and the Skills for Success are made through Goal 1 - Skills and Processes. Shaded blocks indicate a match.

SCIENCE EXPECTATIONS	SKILLS FOR SUCCESS EXPECTATIONS																
	1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																	
1.2																	
1.3																	
1.4																	
1.5																	
1.6																	
1.7																	

RECOMMENDATIONS ABOUT ITEMS AND MATERIALS IN TESTING ROOM:

The Committee recommends the following:

1. not available: dictionaries (of any type), calculators, maps, word lists (of any type), and analytical scoring rubrics.
2. available: the posting of “generic steps in the scientific method” and “general scoring tools”.
3. no more than 5% of the items be of the “negative stem” type. This includes but is not limited to items containing “not”, “never”, “EXCEPT”, “less”, and “least”. Also, NO items of the Roman format appear on the test.
4. 50% of the test can be item sets.
5. at least 25% of the items should have visuals associated with them.

ITEM TYPE DISTRIBUTION:

Brief Constructed Response (BCR):

Six BCRs for a total of approximately 30 minutes. One or more items will be based on the laboratory component.

Extended Constructed Response (ECR)

Three ECRs for a total of approximately 45 minutes: One ECR should be a product of student work (e.g., labeled diagram, graph, or chart) and one should be from the laboratory component.

Selected Response (SR)

Ninety SR questions for a total of approximately 90 minutes.

The SR questions will include discrete questions, classification-set questions, laboratory-set, or technical-passage set questions. Directions for each type of SR will be provided.

MULTICULTURAL CONCERNS:

There should be minority topics or references in 15% of test materials that contain people.

There should be gender balance in materials and references that contain people.

TEST SUBSCORES:

The following subscores are recommended:

Goal 3, Expectation 1 and 2

Goal 3, Expectation 3 and 4

Goal 3, Expectation 5 and 6

EQUIPMENT FOR CLASSROOM LABORATORY WORK:

Microscope, computers, stereoscopes, data gathering software and hardware, electronic balance, thermometer, pH meter, environmental study kit, VCR materials, water bath, hot plate, graphing calculator, biotechnology-related equipment (e.g., gel-electrophoresis boxes, micropipetters)

TOPICS AND CONCEPTS FOR SUGGESTED CLASSROOM LABORATORY EXERCISES:

E1

- I1 Food analysis
Vitamin C analysis (quantity and temperature effects)
- I2 Selectively permeable membrane lab
Enzyme activity lab (temperature, pH)
- I3 Photosynthesis rate lab
Aerobic respiration rate lab (plant or animal)

E2

- I1 Mitosis/meiosis lab
Protein synthesis simulation
Plant and animal physiology (heart rate, breathing, transpiration)
- I2 Environmental/biotic/abiotic requirements
Toxins (*Daphnia*)

E3

- I1 Inheritance patterns - (living organisms/simulation)
Meiosis/mutations
- I2 Inheritance patterns
- I3 Modeling nucleic acid structure/function

E4

- I1 Natural selection/selection simulation
Animal/plant variation
- I2 Classification labs/keying labs
DNA sequencing lab/biotechnology

E5

- I1 Symbiotic relationship lab
- I2 Succession simulation lab
- I3 "Ecosystem" study

CHEMISTRY

Chemistry	Percent of Testing Time*
Goal 4	
E1	7
E2	33
E3	18
E4	32
E5 & 6	10
TOTAL	100

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

Chemistry Skills and Processes

Chemistry	Percent of Testing Time*
Goal 1	
E1	5
E2	25
E4	40
E5	15
E6	7
E7	8
TOTAL	100

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

* Tie to Goal 4, Expectations 5 & 6

Chemistry
Links Between Goals 1 and 4

Goal 4	E1 Openness & Skepticism	E2 Exper. Appr.	E3 Select Instruments	E4 Data Analysis	E5 Communi- cations	E6 Math	E7 Interdis- ciplinary
E1 I1		X	X				
I2		X	X	X		X	
I3		X			X		
E2 I1	X	X			X		
I2	X			X			
I3	X	X			X		
I4		X	X	X		X	
I5		X			X		X
I6							
E3 I1		X	X	X			
I2		X	X	X	X		
I3	X	X	X	X		X	
I4		X	X	X	X		
E4 I1						X	X
I2	X					X	
I3							X
I4		X	X	X			
I5							X
I6		X	X	X	X		
E5 I1							X
I2							X
I3				X			
I4	X			X			X
I5	X			X	X		
E6 I1							X
I2						X	X
I3	X						X

Chemistry Link to Skills for Success

The links of Goal 4 to Skills for Success are made through the links of Goal 1 to Skills for Success. Shaded blocks indicate a match.

SCIENCE EXPECTATIONS	SKILLS FOR SUCCESS EXPECTATIONS																
	1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																	
1.2																	
1.3																	
1.4																	
1.5																	
1.6																	
1.7																	

DISTRIBUTION OF ITEM TYPES:

Brief constructed response (BCR):

Nine BCRs for a total of approximately 45 minutes. Some of these should be tied to recommended lab concepts in the list included with these specifications.

Extended constructed response (ECR):

Two ECRs for a total of approximately 30 minutes. At least one ECR should involve constructing a graph from data and will involve some analysis of that graph.

Selected response (SR):

Ninety SR items for a total of approximately 90 minutes. The SR items may include discrete questions, classification-set questions, laboratory-set questions or technical-passage set items. Directions for each type of SR will be provided.

MULTICULTURAL CONCERNS:

There should be minority topics or references in 15% of test materials that contain people. There should be gender balance in materials and references that contain people. The performance of various subgroups on different types of questions should be investigated. The test items should be written to avoid sophisticated vocabulary or linguistic constructions that are not necessary for content assessment.

EQUIPMENT NEEDED FOR TEST:

Calculator: A four-function calculator is needed. All variety of scientific, programmable, and graphing calculators are allowable, but QWERTY keyboards are *not* allowable. There are no content concerns that would require that memories be cleared before the test, but to maintain security of the test questions calculators with alpha-numeric capabilities should be cleared after the test. The school should have some calculators available for students who do not have their own.

A periodic table should be provided with symbol, mass number, and atomic number only. There should be NO wall-chart periodic table in the testing room.

Room displays or personal aids that contain any information that could be helpful on the test are NOT allowed. This includes both specific chemistry information as well as general information (e.g. a sample graph with all the required components labeled).

Dictionaries of any type will not be allowed.

TEST SUBSCORES:

Suggested subscores are:

1. Expectation 2
2. Expectation 3
3. Expectation 4

Expectation 1 is embedded in each of the above expectations. Remediation for any of the above should include review of the concepts in expectation 1.

TOPICS AND CONCEPTS FOR SUGGESTED CLASSROOM LABORATORY EXERCISES AND ASSOCIATED EQUIPMENT NEEDED

1. Density -- Construct and interpret a graph of mass/volume data
 - graduated cylinder
 - balance
 - graphing calculator
2. Classification of matter -- Demonstrate how matter may be identified and classified in various ways based on common properties
3. Flame test -- Recognize that metallic salts can be identified using a flame test
 - burner
 - spectroscope

4. Identifying bond type -- Recognize and identify bond type of various substances based on
 - solubility, conductivity, and structure
 - conductivity tester
 - atomic model kit
5. Specific heat -- Experimentally determine one or more of: specific heat of a solution or metal sample, or heat of fusion of a sample
 - calorimeter
 - balance
 - thermometer/CBL & temperature probe
6. Conservation of mass -- Verifying conservation of mass via experimentation - reactions in a closed system versus an open system
7. Gas laws -- Observe qualitative relationship between pressure, volume, and temperature of a gas Boyle's law apparatus/syringe and cup
8. Ionic solids -- Classify reactions and write corresponding equations based on observable changes - double replacement reactions
9. Percent composition -- Gather and interpret data
10. pH titration, indicators -- Investigate properties of acids, bases, and salts
 - various household products
 - pH meter/CBL with pH probe
 - burettes
 - variety of indicators
11. Freezing point depression -- Investigate the effects of a solute on the freezing and boiling points of a solution
 - thermometer/CBL & temperature probe
12. Probability -- Become familiar with basic concepts of probability

PHYSICS

Physics	Percent of Testing Time*
Goal 5	
E1	42
E2	21
E3	5
E4	24
E5	8
E6**	0
E7**	0
TOTAL	100%

* The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

** These are testable via links similar to those for Goal 1. See table on next page.

Physics Skills and Processes

PHYSICS	Percent of Testing Time*
Goal 1	
E1	5
E2	10
E3	0
E4	20
E5	40
E6	50
E7	15 -Fulfilled by putting content problems in real-world settings
TOTAL	100

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

Physics
Links Between Goals 1 and 5

Goal 5	E1 Openness & Skepticism	E2 Experimental Approach	E3 Select Instruments	E4 Data Analysis	E5 Communi- cations	E6 Math	E7 Interdis- ciplinary	Goal 5 E6	Goal 5 E7
E1 I1	X	X	X	X	X	X	X	X	X
I2		X	X	X	X	X	X	X	X
I3	X	X	X	X	X	X	X	X	X
I4	X	X	X	X	X	X	X	X	X
I5	X	X	X	X	X	X	X	X	X
E2 I1		X	X	X	X	X	X	X	X
I2		X	X	X	X	X	X	X	X
I3		X	X		X		X	X	X
I4		X	X	X	X		X		X
E3 I1		X		X	X	X	X	X	X
E4 I1&2		X					X		X
I3		X	X	X	X	X	X	X	X
I4	X	X	X	X	X	X	X	X	X
E5 I1	X				X	X	X		X
I2		X		X	X	X	X	X	X
E6 I1									
I2	X			X	X		X	---	---
I3	X			X	X		X	---	---
I4	X			X	X	X	X	---	---
E7 I1							X	---	---
I2				X			X	---	---
I3						X	X	---	---
								---	---

Physics Link to Skills for Success

The links of Goal 5 to Skills for Success are made through the links of Goal 1 to Skills for Success. Shaded blocks indicate a match.

SCIENCE EXPECTATIONS	SKILLS FOR SUCCESS EXPECTATIONS																
	1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	5.3	5.4
1.1																	
1.2																	
1.3																	
1.4																	
1.5																	
1.6																	
1.7																	

DISTRIBUTION OF ITEM TYPES:

Constructed response items

The constructed response items can require mathematical problem-solving, essays, explanations, or any combination of these.

Brief constructed response (BCR):

Six BCRs for a total of approximately 30 minutes. Some of these will be laboratory related, either in setting or skills required.

Extended constructed response (ECR):

Three ECRs for a total of approximately 45 minutes. Some of these may involve constructing a graph from data and involving analysis of that graph.

Selected response (SR) items

Ninety SR questions for a total of approximately 90 minutes. The SR questions will include discrete questions, questions based on a technical passage, classification-set questions and laboratory-set questions. Directions for each type of SR will be provided.

MULTICULTURAL CONCERNS:

There should be a minority topic or reference in 15% of test materials that contain people. There should be gender balance in materials and references that contain people. The performance of various subgroups on different types of questions should be investigated.

EQUIPMENT NEEDED FOR TEST:

Calculator: At least a scientific calculator is needed (one with trigonometric functions, square roots, and exponential notation). All variety of programmable and graphing calculators are allowable, but QWERTY keyboards are *not* allowable. There are no content concerns that would require that memories be cleared. The school should have some calculators available for students who do not have their own.

A table of constants, a list defining symbols that will be used on the test, and a list of basic equations will be provided with the test.

There should be *no* material posted in the testing room that contains information beyond what is provided with the test. It is recommended that tests be administered in rooms not used for the subject (e.g. science tests administered in English rooms)

Dictionaries of any type will not be permitted.

TEST SUBSCORES:

Suggested subscores are:

1. Expectation 1
2. Expectations 2 and 3
3. Expectations 4 and 5

TOPICS AND CONCEPTS FOR SUGGESTED CLASSROOM LABORATORY EXERCISES AND ASSOCIATED EQUIPMENT NEEDED:

The list of equipment is cumulative. A laboratory exercise may require some equipment included under a previous lab in the list.

Kinematics -- Relationships between displacement, velocity, and acceleration for situations of constant acceleration

- timing device (e.g., stopwatch, ticker tape)
- moving object (carts, etc.)
- distance measure (e.g., ruler)

Newton's laws -- Application of Newton's laws in static and dynamic situations

- force measurement device (e.g., scale) or calibrated weights and mass measurement device

Force addition -- Graphical and analytical addition of vectors.

- protractor

Work and Energy -- Understanding when conservation of mechanical energy is and is not applicable

- inclined plane

Momentum -- Conservation of momentum, action and reaction forces

- momentum carts

Simple circuits -- Behavior of series and parallel resistor circuits, loop and junction circuit rules

- voltmeters
- ammeters
- power supply
- resistors/light bulbs
- wires
- switches

Electromagnetism -- Concept of magnetic field, production of field by permanent magnets and currents

- metal objects (e.g. nails)
- device to detect a magnetic field (e.g., compass, iron filings)
- permanent magnets

Wave behavior -- Wave properties, superposition (In some cases demonstration may be more manageable than a lab for this topic)

- source of waves (generator, laser, tuning fork)
- wave visualizer (ripple tank, Slinky, etc.)

Diffraction, Interference -- Qualitative and quantitative understanding of diffraction and interference patterns

- monochromatic light source
- interference slits or diffraction grating

Refraction -- Snell's law

- glass plate
- light source or objects (e.g. pins)

Resonance -- Concept of resonance (small input yields large output via standing wave), relationship of frequency, wavelength, and speed

- tuning fork or wave generator
- cylinder (for adjustable water column) or string

Mirrors and Lenses -- Production of real and virtual images, lens equation

- mirrors
- card with pinhole
- lenses
- extended light source
- screen

Nuclear decay simulation -- Half-life, random individual events versus group predictability

- "coins"

The list of equipment was developed with the intention of including the most affordable items that can be employed. Skills for Success will require phase-in of more sophisticated equipment (e.g., lasers, computers, CBL) to meet the technology goals.

Specifications for the Maryland HSAs in Social Studies

“PACKET” OF INSTRUCTIONAL MATERIALS FOR TEACHERS OF SOCIAL STUDIES

The Social Studies Content Specifications Committee strongly recommends that all social studies teachers be provided with a packet of materials to aid instruction throughout the year in all Government, US History, and World History courses. This packet should include:

- List of “Recommendations to Teachers” and suggestions on how to use the materials in the packet.
- “How to...” tips regarding what students should know about and how teachers should teach certain key skills necessary in social studies. For example, something the content committee agrees should be in every “packet” is document analysis skills (i.e., identifying point of view, revealing bias, distinguishing fact from opinion, defining the main idea, uncovering supporting detail). Other skills that could be included: decision-making and problem-solving, interpretation of representational art, and the analysis of primary historical documents.
- Model instructional activities for teachers to use in order to help prepare students for the assessment. Activities should be linked explicitly to the Core Learning Goals and Skills for Success and might include, but are not limited to: cartoon interpretation, document analysis, and interpretation of charts, maps, tables, and graphs.
- Several samples of each item type (ECR, BCR, and SR).
- Samples of item “sets” (see description above).
- Sample of a document-based ECR.
- List of content-specific vocabulary.
- Sample scoring rubrics for BCR’s and ECR’s.

UNITED STATES HISTORY

	% of Testing Time* (165 min.)	Link to Skills for Success
Goal 1	20.3%	
E1		2.2, 3.1, 3.2
E2		2.2, 2.3, 3.1, 3.2
Goal 2	39.1%	
E1		2.1, 2.2, 2.3, 2.4, 3.1, 3.2
E2		2.2, 2.3, 2.4, 3.1, 3.2
Goal 3	20.3%	
E1		2.1, 2.2, 3.1, 3.2, 4.3
E2		2.1, 2.2, 2.3, 3.1, 3.2, 4.3
Goal 4	20.3%	
E1		2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 4.3
TOTAL	100%	

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

DISTRIBUTION AND DEFINITION OF ITEM TYPES:

Brief Constructed Response (BCR)

Eight BCRs for a total of approximately 40 minutes. A BCR is a response that allows for the brief development of an idea. These could be two-part questions, could test any aspect of the test specifications, and might include entries on a chart, lists, timelines, or other graphic representations.

Extended Constructed Response (ECR)

One ECR of approximately 30 minutes. An ECR is an appropriate and extended written response that includes planning ("pre-writing") and may be a document-based exercise.

Selected Response (SR)

Sixty selected-response questions for a total of approximately 90 minutes. Some SR items will be linked in “sets” based on a stimulus that may also include a BCR. The content committee recognizes that the number of selected response questions indicated for this test is inconsistent with the original working assumptions and current recommended technical specifications. The number of selected-response questions may need to be increased to 90 or more as a result of pilot and no-fault tests in order to meet psychometric requirements. In addition, more constrained, less open-ended constructed response items and scoring tools may be advisable in order to ensure high rater agreement rates.

TEST SUBSCORES:

Recommended subscores are:

1. Goal 1
2. Goal 2
3. Goal 3
4. Goal 4

MULTICULTURAL CONCERNS:

There should be minority topics or references in at least 15% of test materials. There should be gender balance in materials and references.

GOVERNMENT

	% of Testing Time* (165 min.)	Link to Skills for Success
Goal 1	59.4%	
E1		2.1, 2.2, 2.3, 2.4, 3.2
E2		2.1, 2.2, 2.3, 2.4, 3.2
Goal 2	17.5%	
E1		2.2, 3.2, 4.3
E2		2.2, 3.2, 4.3
Goal 3	8.7%	
E1		2.2, 3.2
Goal 4	14.4%	5.4
E1		2.2, 3.2
TOTAL	100%	

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

DISTRIBUTION AND DEFINITION OF ITEM TYPES:

Brief Constructed Response (BCR)

Eight BCR for a total of approximately 40 minutes. A BCR is a response that allows for the brief development of an idea. These could be two-part questions, could test any aspect of the test specifications, and might include entries on a chart, lists, timelines or other graphic representations.

Extended Constructed Response (ECR)

One ECR of approximately 30 minutes. An ECR is an appropriate and extended written response that includes planning ("pre-writing") and may be a document-based exercise.

Selected Response (SR)

60 selected-response questions for a total of approximately 90 minutes. Some SR items will be linked in “sets” based on a stimulus that may also include a BCR. The content committee recognizes that the number of selected-response questions indicated for this test is inconsistent with the original working assumptions and current recommended technical specifications. The number of selected-response questions may need to be increased to 90 or more as a result of pilot and no-fault tests in order to meet psychometric requirements. In addition, more constrained, less open-ended constructed response items and scoring tools may be advisable in order to ensure high rater agreement rates.

TEST SUBSCORES:

Recommended subscores are:

1. Goal 1
2. Goal 2
3. Goals 3 and 4

MULTICULTURAL CONCERNS

There should be minority topics or references in at least 15% of test materials.

There should be gender balance in materials and references.

WORLD HISTORY

	% of Testing Time* (165 min.)	Link to Skills for Success
Goal 1	21.7%	
E1		2.2, 3.2, 4.3
Goal 2	40.5%	
E1		2.1, 2.2, 2.3, 2.4, 3.2
E2		2.2, 3.2
E3		2.2, 2.3, 3.2
Goal 3	18.9%	
E1		2.2, 2.4, 3.2, 4.3
E2		2.2, 3.2
E3		2.2, 3.2, 4.3
Goal 4	18.9%	
E1		2.2, 3.2, 4.3
TOTAL	100%	

*The percentage of testing time is not necessarily distributed evenly across all indicators within an expectation.

DISTRIBUTION AND DEFINITION OF ITEM TYPES:

Brief Constructed Response (BCR)

Eight BCRs for a total of approximately 40 minutes. A BCR is a response that allows for the development of an idea (taking more or less 5 minutes). These could be two-part questions, could test any aspect of the test specifications, and might include entries on a chart, lists, timelines, or other graphic representations.

Extended Constructed Response (ECR)

One ECR of approximately 30 minutes. An ECR is an appropriate and extended written response that includes planning ("pre-writing") and may be a document-based exercise.

Selected Response (SR)

Sixty selected-response questions for a total of approximately 90 minutes. Some SR items will be linked in "sets" based on a stimulus that may also include a BCR. The content committee recognizes that the number of selected response questions indicated

for this test is inconsistent with the original working assumptions and current recommended technical specifications. The number of selected-response questions may need to be increased to 90 or more as a result of pilot and no-fault tests in order to meet psychometric requirements. In addition, more constrained, less open-ended constructed response items and scoring tools may be advisable in order to ensure high rater agreement rates.

TEST SUBSCORES:

Recommended subscores are:

1. Goal 1
2. Goal 2
3. Goal 3
4. Goal 4

MULTICULTURAL CONCERNS

There should be minority topics or references in at least 15% of test materials. There should be gender balance in materials and references.

Appendix D

Illustrative Items

The illustrative items included in this section are intended to represent the major types of questions that will be included in the HSA examinations and scoring guides. These illustrations should not be considered comprehensive, or samples of the actual items that will appear in the HSA examinations. Actual items will be developed in the test development process and will be released for public information.

The items used in HSA tests will be scored using generic rubrics. These generic rubrics, which represent the standards that will be used in scoring brief constructed and extended constructed response questions in each test will be published in advance of the HSA examinations. The rubrics will evolve during the test development process and will detail the features of each point used in scoring. The rubrics will be useful for teachers who can see how their students' work will be evaluated. At the same time, they will provide information to the students that is intended to help them understand the characteristics that will be assessed in their constructed response answers.

Each illustrative question contains information about the test (as indicated by the test heading or question heading), followed by the links to the Core Learning Goals and Skills for Success.

ENGLISH

Questions 1 - 6 refer to the poems below.

Dream Deferred

by Langston Hughes

Harlem

What happens to a dream deferred?

Does it dry up
like a raisin in the sun?
Or fester like a sore--
And then run?
Does it stink like rotten meat?
Or crust and sugar over--
like a syrupy sweet?

(5)

Maybe it just sags
like a heavy load.

(10)

Or does it explode?

Alabama Centennial

by Naomi Long Madgett

They said, "Wait." Well, I waited.
For a hundred years I waited
In cotton fields, kitchens, balconies,
In bread lines, at back doors, on chain gangs,
In stinking "colored" toilets (5)
And crowded ghettos,
Outside of schools and voting booths.
And some said, "Later."
And some said, "Never!"

Then a new wind blew, and a new voice (10)
Rode its wings and quiet urgency
Strong, determined, sure.
"No," it said. "Not 'never,' not 'later,'
Not even 'soon.'
Now. (15)
Walk!"

And other voices echoed the freedom words,
"Walk together, children, don't get weary,"
Whispered them, sang them, prayed them, shouted them.
"Walk!" (20)
And I walked in streets in Montgomery⁽¹⁾
Until a link in the chain of patient acquiescence broke.

Then again: Sit down!
And I sat down at the counters of Greensboro.⁽²⁾
Ride! And I rode the bus for freedom (25)
Kneel! And I went down on my knees in prayer and faith.
March! And I'll march until the last chain falls
Singing, "We shall overcome."

Not all the dogs and hoses in Birmingham⁽³⁾
Nor all the clubs and guns in Selma⁽⁴⁾ (30)
Can turn this tide
Not all the jails can hold these young black faces
From their destiny of manhood.
Of equality, of dignity,
Of the American Dream (35)
A hundred years past due.
Now!

(1)Montgomery: Capital of Alabama: in 1955 Martin Luther King, Jr., led a boycott there that ended racial segregation on buses.

(2)Greensboro: The largest city in North Carolina: in 1960 four African American students sat at a restricted lunch counter in Greensboro to protest racial segregation. This action prompted a wave of sit-in demonstrations throughout the South.

(3)Birmingham: In 1963 racial tensions in Birmingham, Alabama, escalated until a bomb exploded in an African American church, killing four girls. As a result, interracial groups organized and began working to prevent future incidents.

(4)Selma: In 1965 King led a five-day march from Selma, Alabama, to Montgomery to protest discrimination in voter registration. Later that year Congress passed the Voting Rights Act, which gave 100,000 African Americans the right to vote.

Selected Response

1. Content CLG: Goal 1, Expectation 2, Indicator 3
Skills for Success: 1.2.3

The word "deferred" (line 2) as used in the poem "Dream Deferred" is closest in meaning to which of the following choices?

- A. Fulfilled
- *B. Delayed
- C. Awakened
- D. Revealed

2. Content CLG: Goal 1, Expectation 2, Indicator 3
Skills for Success: 1.2.3

Which organizational element contributes most to a sense of tension and climax in "Dream Deferred?"

- *A. Six questions with a statement before the final question
- B. A sub-heading which identifies the setting
- C. Four stanzas of varying length
- D. The repetition of the word "it"

3. Content CLG: Goal 1, Expectation 2, Indicator 1
Skills for Success: 1.2.3

Which word below best identifies the climax of "Dream Deferred?"

- A. Stink
- B. Crust
- C. Say
- *D. Explode

Brief Constructed Response

4. Content CLG: Goal 1, Expectation 2, Indicator 2
Skills for Success: 1.2.3

Choose one of the five similes or comparisons to a deferred dream in Langston Hughes' poem and copy the simile in your answer booklet. After rereading the simile, explain what you think the speaker of the poem is trying to suggest by using this comparison. Try to refer to specific words from the simile in your explanation.

5. Content CLG: Goal 1, Expectation 2, Indicator 2
Skills for Success: 1.2.3

In the poem "Alabama Centennial" the phrase "patient acquiescence" conveys how some African Americans waited uncomplainingly for civil rights. Explain the meaning and the appropriateness of the metaphor in line 22 which helps the reader to envision the end of such "patient acquiescence."

6. Content CLG: Goal 1, Expectation 2, Indicator 2
Skills for Success: 1.2.3

In "Alabama Centennial," a series of verbs illustrates the progression of the American civil rights movement. Select one of the verbs Madgett emphasizes and explain how you know it is important in the poem.

Questions 7 - 10 refer to the excerpt below from *Talent* by Annie Dillard.

There is no such thing as talent. If there are any inborn, God-given gifts, they are in the precocious fields of music, mathematics, and chess; if you have such a gift, you know it by now. All the rest of us, in all the other fields, are not talented. We all start out dull and weary and uninspired. Apart from a few like Mozart, there never have been any great and accomplished little children in the world. Genius is the product of education.... (5)

It's hard work, doing something with your life. The very thought of hard work makes me queasy. I'd rather die in peace. Here we are, all equal and alike and none of us much to write home about--and some people choose to make themselves into physicists or thinkers or major-league pitchers, knowing perfectly well that it will be nothing but hard work. But I want to tell you that it's not as bad as it sounds. Doing something does not require discipline; it creates its own discipline. (10)

People often ask me if I discipline myself to write, if I work a certain number of hours a day on a schedule. They ask this question with envy in their voices and awe on their faces and a sense of alienation all over them, as if they were addressing an armored tank or a talking giraffe or Niagara Falls. We all want to believe that other people are natural wonders; it gets us off the hook. (15)

Now, it happens that when I wrote my first book of prose, I worked an hour or two a day for a while, and then in the last two months, I got excited and worked very hard, for many hours a day. People can lift cars when they want to. People can recite the Koran, too and run in marathons. These things aren't ways of life; they are merely possibilities for everyone on certain occasions of life. You don't lift cars around the clock or write books every year. But when you do, it's not so hard. It's not superhuman. It's very human. You do it for love. You do it for love and respect for the task itself.... (20)

Of course it has to be done. And something has to be done with your life too: something specific, something human. But don't wait around to be hit by love. Don't wait for anything. Learn something first. Then while you are getting to know it, you will get to love it and that love will direct you in what to do. (25)

Dillard, Annie. "Is There Really such a Thing as Talent" (titled "Talent"). Blanche C. Gregory, Inc., 1979.

Brief Constructed Response

7. English 1
Content CLG: Goal 1, Expectation 3, Indicator 3
Skills for Success: 1.2.3
- Summarize Dillard's ideas on "doing something with your life" (line 7). Use text support in your response.
8. English 1
Content CLG: Goal 1, Expectation 3, Indicator 5
Skills for Success: 1.2.3
- Themes in literature often reflect experiences common to people across time and cultures. One such theme is that ordinary people are capable of superhuman effort. Cite and explain an example from "Talent" that supports this theme.
9. English 1
Content CLG: Goal 3, Expectation 2, Indicator 1
Skills for Success:
- Annie Dillard develops her ideas by repeating and varying several every-day words and phrases. Choose *one* of the underlined examples below; first, define it as it is commonly used, and then explain Dillard's particular variations of meaning.
- talent hard work discipline love
10. English 1
Content CLG: Goal 1, Expectation 1, Indicator 2
Skills for Success: 1.2.3
- Study lines 17-18. After rereading the lines, generate a list of questions that the sentence brings to mind. What questions could you ask that could help you to interpret Dillard's text as a whole more thoroughly?

Extended Constructed Response

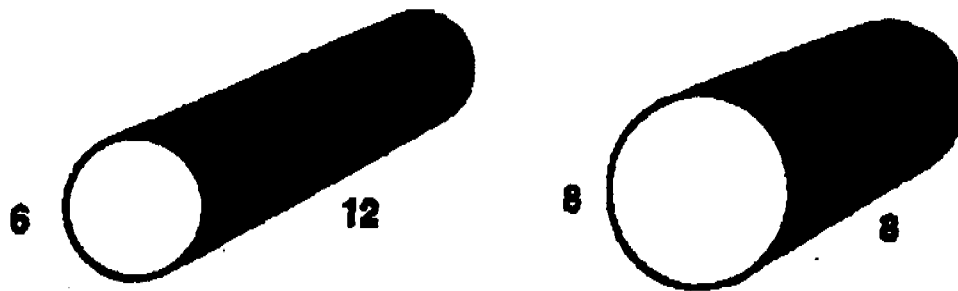
Note: Question 11. does not refer to a specific text.

11. Content CLG: Goal 2, Expectation 1, Indicator 3
Skills for Success:
- Someone once said that "Genius is the product of education." What does this statement mean to you? Using your personal, academic, and/or literary experiences for support, write an essay in which you agree or disagree with the statement that "Genius is the product of education." Remember to extend and elaborate your ideas.

MATHEMATICS

Selected Response

1. Content CLG: Goal 2, Expectation 3, Indicator 2
Skills for Success: 2.1.1, 2.2.1, 2.2.2, 2.2.4, 2.4.1, 2.4.2, 2.4.3, 2.4.4



The figure above shows two open-ended cylindrical pipes that are made from sheet metal. One has diameter 6 inches and length 12 inches and the other has diameter 8 inches and length 8 inches. What is the surface area of the one that can be made with less sheet metal?

- *A. 64π square inches
- B. 72π square inches
- C. 108π square inches
- D. 128π square inches

Machine Scorable Student-Produced Response

Directions

Questions 2. and 3. require you to solve the problem and enter your answer by marking the ovals in the special grid, as shown in the examples below.

Answer: $\frac{7}{12}$ or 7/12

Write answer in boxes. →

7	/	1	2
○	○	○	○
1	1	0	1
2	2	2	0
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9

Grid in result. →

← Fraction line

Answer: 2.5

2	.	5
○	○	○
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9	9	9

← Decimal point

Answer: 201
Either position is correct.

2	0	1
○	○	○
1	1	1
2	2	2
3	3	3
4	4	4

2	0	1
○	○	○
1	1	1
2	2	2
3	3	3
4	4	4

Note: You may start your answers in any column, space permitting. Columns not needed should be left blank.

- Mark no more than one oval in any column.
- Because the answer sheet will be machine-scored, you will receive credit only if the ovals are filled in correctly.
- Although not required, it is suggested that you write your answer in the boxes at the top of the columns to help you fill in the ovals accurately.
- Some problems may have more than one correct answer. In such cases, grid only one answer.
- No question has a negative answer.
- **Mixed numbers** such as $2\frac{1}{2}$ must be gridded as 2.5 or 5/2. (If $\frac{2\frac{1}{2}}{2}$ is gridded, it will be interpreted as $\frac{21}{2}$, not $2\frac{1}{2}$.)

- **Decimal Accuracy:** If you obtain a decimal answer, enter the most accurate value the grid will accommodate. For example, if you obtain an answer such as 0.6666..., you should record the result as .666 or .667. Less accurate values such as .66 or .67 are not acceptable.

Acceptable ways to grid $\frac{2}{3} = .6666...$

2	/	3
○	○	○
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6

.	6	6	6
○	○	○	○
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6

.	6	6	7
○	○	○	○
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6

2. Content CLG: Goal 1, Expectation 2, Indicator 5
Skills for Success: 2.2.2, 2.2.3, 2.4.1, 2.4.2, 2.4.3, 2.4.4

	New				New		
	Rap	Rock	Age		Rap	Rock	Age
<u>CD's</u>	401	562	212	<u>CD's</u>	341	562	212
Cassettes	374	609	347	Cassettes	374	527	312

The matrix on the left above gives data about the inventory of CD's and cassettes at Power Records store at the beginning of the day on Tuesday. The matrix on the right above gives data about the inventory at the beginning of the day on Wednesday. If the store did not get any more CD's and cassettes in stock on Tuesday, how many rock cassettes did Power Records sell during the day on Tuesday?

3. Content CLG: Goal 2, Expectation 3, Indicator 1
Skills for Success: 2.1.1, 2.2.1, 2.2.2, 2.2.3, 2.4.1, 2.4.2, 2.4.3, 2.4.4

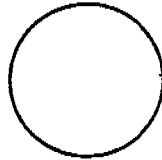
At a certain time of day, Lisa's shadow is 8 feet long and the shadow of the tree next to her is 64 feet long. If Lisa is 5 feet 6 inches tall, how many feet tall is the tree?

Brief Constructed Response

4. Content CLG: Goal 2, Expectation 3, Indicator 2
Skills for Success: 2.1.1, 2.2.1, 2.2.2, 2.2.3, 2.4.1, 2.4.2, 2.4.3, 2.4.4

Diameter = 10 cm

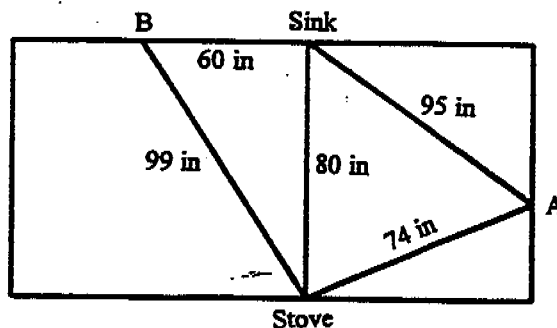
8.5 cm



7.5 cm

The figure above shows the exposed surface of a circular solar cell and a rectangular solar cell. If solar cells generate approximately 18 milliamps of electrical current for each square centimeter of solar cell exposed to direct sunlight, which cell will generate more milliamps of electricity? How much more? Show how you arrived at your answer.

5. Content CLG: Goal 2, Expectation 3, Indicator 2
Skills for Success: 2.1.1, 2.2.1, 2.2.2, 2.2.3, 2.4.1, 2.4.2, 2.4.3, 2.4.4



The work triangle of a kitchen is determined by the location of the sink, refrigerator and stove. Because of plumbing and electrical requirements, the stove and sink in the diagram of the kitchen above must be placed as shown. Points A and B show possible locations for the refrigerator. If the perimeter of the work triangle should not exceed 21 feet, use mathematics to show why both A and B are suitable locations for the refrigerator.

6. Content CLG: Goal 3, Expectation 1, Indicator 2
Skills for Success: 2.2.1, 2.2.2, 2.2.3, 2.4.1, 2.4.2, 2.4.3, 2.4.4

Model	Mileage (Miles per gallon)
A	28
B	28
C	31
D	32
E	35

The table above shows the gas mileage for five different car models that are sold by a car manufacturer. The manufacturer wants to advertise this group of cars by publishing a single number for gas mileage that is representative of these five models. Which of the following, the mean, the median, or the mode of the gas mileages in the table, is the greatest? Use mathematics to show your work to justify your answer.

Extended Constructed Response

7. Content CLG: Goal 1, Expectation 1, Indicator 2; Goal 1, Expectation 2, Indicator 1
Skills for Success: 2.2.1, 2.2.2, 2.2.3, 2.4.3, 2.4.4, 3.1.3, 4.1.4, 3.2.5

You and a few of your friends are forming a video game club. Members of the club will be able to rent video games for a certain fee. You and your friends, have decided to offer two different membership payment plans.

The two payment plans are as follows:

Plan A: \$5 initiation fee plus \$1.50 per video game rental
Plan B: No initiation fee, \$2 per video game rental

- a) Represent the cost of renting video games for each plan. You may use an equation, table, or graph. (A table or graph should illustrate the cost of renting 1 through 12 games.)
- b) Each of these plans may be beneficial for different members. Use the language of mathematics and your equation, table, or graph to explain when each plan is better. Give an example of a situation in which plan A is better and give another example of a situation in which plan B is better.

SCIENCE

Selected Response

1. Earth and Space Science
Content CLG: Goal 1, Expectation 7, Indicator ; Goal 2, Expectation 2, Indicator 1;
Goal 2, Expectation 6, Indicator 2
Skills for Success:

The Big Bang theory of the formation and expansion of the universe is supported by observation of which of the following?

- A. Light from stars experiencing a Doppler shift toward the blue-violet
- *B. Light from stars experiencing a Doppler shift toward the red
- C. Asteroids in the Solar system
- D. The solar wind

Brief Constructed Response

2. Earth and Space Science
Content CLG: Goal 1, Expectation 5, Indicator 1; Goal 2, Expectation 2, Indicator 1
Skills for Success: 2.2, 2.4, 3.1, 3.2

As an object orbits the Sun its speed changes. Describe, *in writing*, the path of an object in orbit around the Sun and indicate reasons for this change in speed. You may include a diagram to help clarify your explanation.

Selected Response

3. Earth and Space Science
Content CLG: Goal 1, Expectation 4, Indicator ; Goal 2, Expectation 4, Indicator 5;
Goal 2, Expectation 1, Indicator 1
Skills for Success:

When trying to locate the epicenter of an earthquake, it is necessary to analyze waves from at least how many seismographic locations?

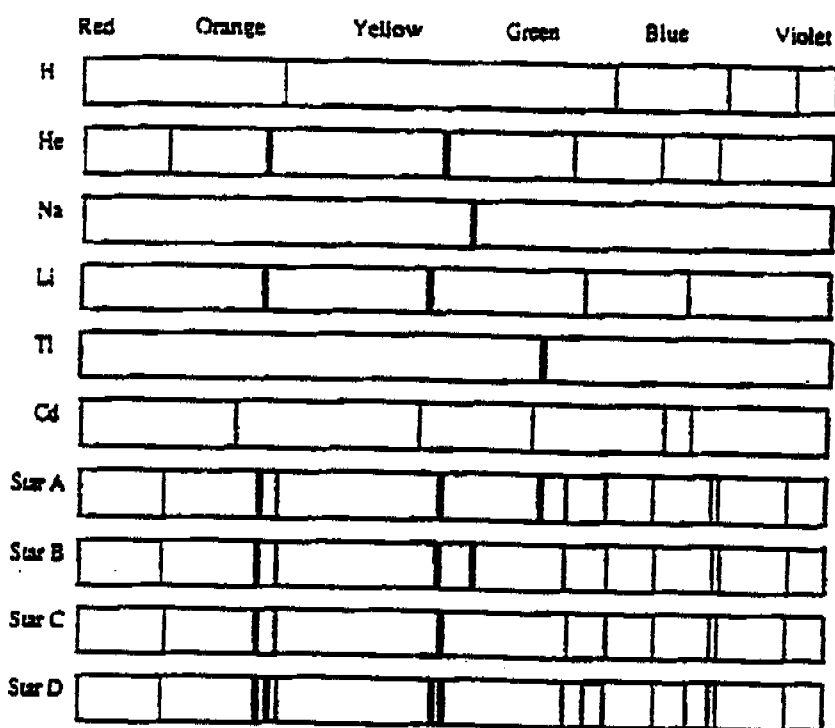
- A. Two
- *B. Three
- C. Four
- D. Five

Selected Response

4. Earth and Space Science
Content CLG: Goal 1, Expectation 4, Indicator 1; Goal 1, Expectation 2, Indicator 5;
Goal 2, Expectation 1, Indicator 2
Skills for Success: 1.3, 2.2, 2.4, 3.2, 4.1

The spectra of known elements can be compared to the spectra of stars to determine the elements present in the stars. Use the spectral lines from FIGURE 1 to answer the question below.

FIGURE 1



Which of the following lists all the elements that can be found in all four stars?

- A. Hydrogen only
- B. Helium only
- *C. Hydrogen and Helium
- D. Hydrogen, Helium, and Sodium

Extended Constructed Response

5. Earth and Space Science

Content CLG: Goal 1, Expectation 5, Indicator 1; Goal 2, Expectation 2, Indicator 1

Skills for Success: 3.1

Suppose that the Moon had a lunar month of 16 days.

- a) Draw a diagram of the Earth-Moon-Sun system which illustrates the moon in its orbit around the Earth. Indicate the position of the new moon, first quarter, full moon, and last quarter by appropriate shading of the figures.
- b) A lunar month begins on June 7 with a new moon. Draw and label the view of the moon as seen from the earth for the new moon, first quarter, full moon and last quarter. Label each diagram with the corresponding date for this lunar month.

Brief Constructed Response

1. Biology

Content CLG: Goal 1, Expectation 5, Indicator 2; Goal 3, Expectation 3, Indicator 1
Skills for Success: 2.4, 3.1, 3.2

In humans, the allele for red-green color blindness is recessive to the allele for normal color vision, and is located on the X chromosome. A color-blind father and an normal-color-vision mother produce a color-blind daughter. Construct a punnett square that illustrates the genotypes of the parents and possible offspring.

Selected Response

2. Biology

Content CLG: Goal 1, Expectation 4, Indicator 5; Goal 3, Expectation 1, Indicator 2;
Goal 3, Expectation 5, Indicator 1
Skills for Success: 1.3, 2.2, 2.4

FIGURE 1

<i>Water Temperature (Degrees Celsius)</i>	<i>Average Number of Gill Movements (per minute)</i>
5	22
10	30
15	35
20	41

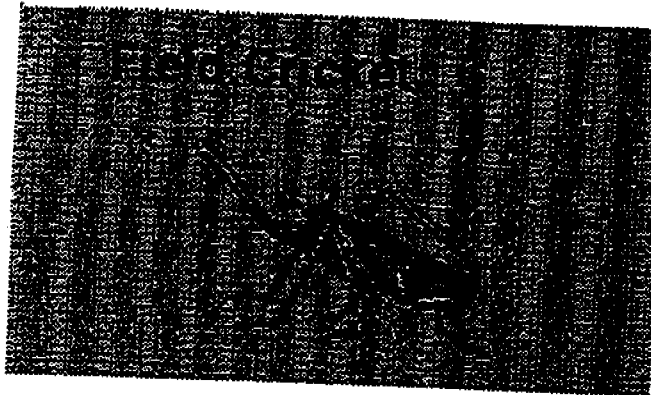
A student obtained the results listed in FIGURE 1 above by performing an investigation with goldfish in four different containers of fresh water. The temperature of each container was carefully controlled and the gill movements of the fish were counted for five minutes and the average results were calculated. Which of the following statements is best supported by these data?

- *A. Gill movements in goldfish increase as temperature increases.
- B. Since goldfish are warm-blooded, gill movements are not related to temperature changes.
- C. If the temperature continued to increase, the goldfish would die.
- D. Gill movements in goldfish decrease with increases in temperature.

Brief Constructed Response

3. **Biology**

Content CLG: Goal 1, Expectation 2, Indicators 1-5; Goal 3, Expectation 2, Indicator 2
Skills for Success: 1.3, 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 4.1, 4.2



After a discussion in Biology class on metabolism in insects, you want to know if the "chirps" from a field cricket are a reliable indicator of air temperature.

- Construct a hypothesis that you could use to guide your investigation.
- Devise an experiment that you might use to gather data to test your hypothesis. Include a numbered procedure and list appropriate equipment from your biology classroom that could be used.
- Describe results that could be obtained and explain how they would support your hypothesis. Describe results that could be obtained and explain how they would not support your hypothesis.

Extended Constructed Response

4. Biology

Content CLG: Goal 1, Expectations 2,4,&5, Indicator ; Goal 3, Expectation 2, Indicator
Skills for Success: 1.3, 2.1, 2.2, 2.3, 3.1, 3.2

The data in Table 1 show the heart rate for an adult during a 10 minute walk on a treadmill. The resting heart rate before the walk was 66 beats per minute.

TABLE 1

Minute	Heart Rate
0	88
1	88
2	100
3	133
4	121
5	107
6	110
7	100
8	105
9	109
10	100
11	87
12	88
13	68
14	67
15	67

INSTRUCTIONS

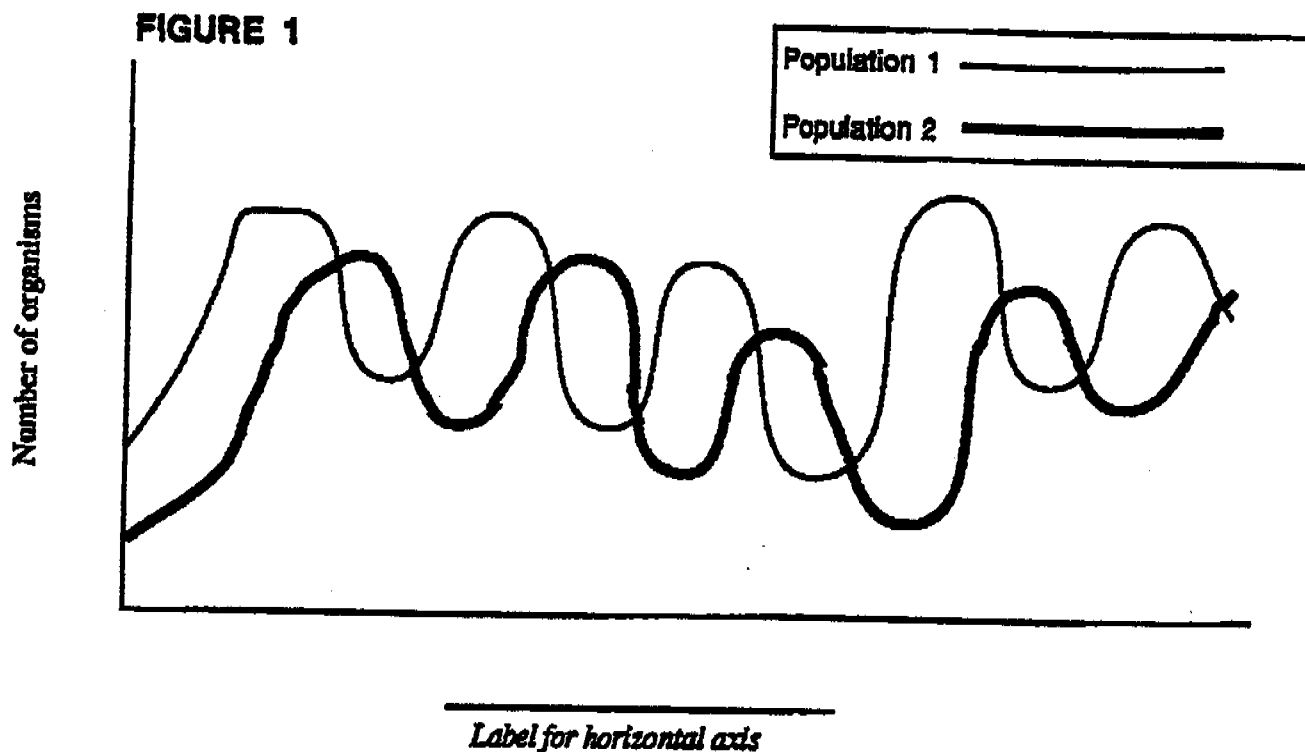
- Plot the points from Table 1 on a graph and connect the points to make a line. Place appropriate labels and units on each axis of the graph.
- Describe the trend in this person's heart rate from time zero to minute 15.
- Describe and provide reasons for physiological changes other than heart rate that could be occurring in this person as a result of this walk on the treadmill.
- State one conclusion that can be made from the data in Table 1.

Brief Constructed Response

5. Biology

Content CLG: Goal 1, Expectation 4, Indicator 1; Goal 1, Expectation 5, Indicator 1;
Goal 3, Expectation 5, Indicator 1

Skills for Success: 1.3, 2.2, 2.4, 3.1, 3.2



Observe the data presented in FIGURE 1 above.

Two populations in nature may act on each other and, at the same time, also be influenced by other factors in the environment. The two lines on the graph represent the number of organisms in two different populations within a community. The numbers of organisms are plotted on the vertical axis ("y" axis) and are represented by two lines, solid and bold.

Place an appropriate label and units on the horizontal axis of the graph.

Write a paragraph to describe the interaction of these two populations of organisms.

- Give examples of the species or type of organisms that may be represented.
- Be sure to include reasons for the graph having this particular shape.
- Discuss how the shape of the graph would change if the number of *producers* in this community would suddenly decrease.

Selected Response

1. Chemistry

Content CLG: Goal 1, Expectation 4, Indicator 1; Goal 4, Expectation 4, Indicator 5
Skills for Success:

A student heats and completely decomposes 100 g of calcium carbonate, which produces 56 g of calcium oxide. The reaction proceeds according to the following equation.



How many grams of carbon dioxide are produced in this reaction?

- *A. 44 g
- B. 100 g
- C. 144 g
- D. 156 g

2. Chemistry

Content CLG: Goal 1, Expectation 7, Indicator 1; Goal 4, Expectation 4, Indicator 4;
Goal 4, Expectation 4, Indicator 5
Skills for Success:

Propane gas (C_3H_8) is used instead of gasoline in some alternative fuel cars. Which of the following represents the combustion reaction that occurs when propane is completely burned?

- A. $\text{C}_3\text{H}_8 + \text{O}_2 \rightarrow \text{C}_3\text{O}_2 + \text{H}_8$
- B. $\text{C}_3\text{H}_8 + \text{O}_2 \rightarrow 3\text{C} + \text{O}_2 + 4\text{H}_2$
- *C. $\text{C}_3\text{H}_8 + 5\text{O}_2 \rightarrow 3\text{CO}_2 + 4\text{H}_2\text{O}$
- D. $3\text{CO}_2 + 4\text{H}_2\text{O} \rightarrow \text{C}_3\text{H}_8 + 5\text{O}_2$

3. Chemistry

Content CLG: Goal 1, Expectation 4, Indicator 3; Goal 4, Expectation 4, Indicator 1
Skills for Success:

A student experimentally determines the molar mass of an unknown compound to be 46.0 grams per mole. Which of the following compounds could the student correctly report as a possible identity of the unknown compound?

- A. MgCl_2
- B. CO_2
- C. C_2H_8
- *D. $\text{C}_2\text{H}_5\text{OH}$

Brief Constructed Response

4. Chemistry

Content CLG: Goal 1, Expectation 4, Indicator 5; Goal 4, Expectation 4, Indicator 1
Skills for Success: 3.1

Calculate the mass (in amu) of a molecule of C_2H_5OH . Show your work and include appropriate units on your answer.

5. Chemistry

Content CLG: Goal 1, Expectation 4, Indicator 5; Goal 4, Expectation 4, Indicator 2;
Goal 4, Expectation 4, Indicator 5
Skills for Success: 3.1

When copper is heated in air, the mass of the resulting product is greater than that of the original copper.

- a) Describe the reaction that occurs when the copper is heated.
- b) Explain why the mass of the product is greater than that of the original copper.

6. Chemistry

Content CLG: Goal 1, Expectation 4, Indicator 5; Goal 4, Expectation 4, Indicator 6;
Goal 4, Expectation 4, Indicator 4
Skills for Success: 3.1

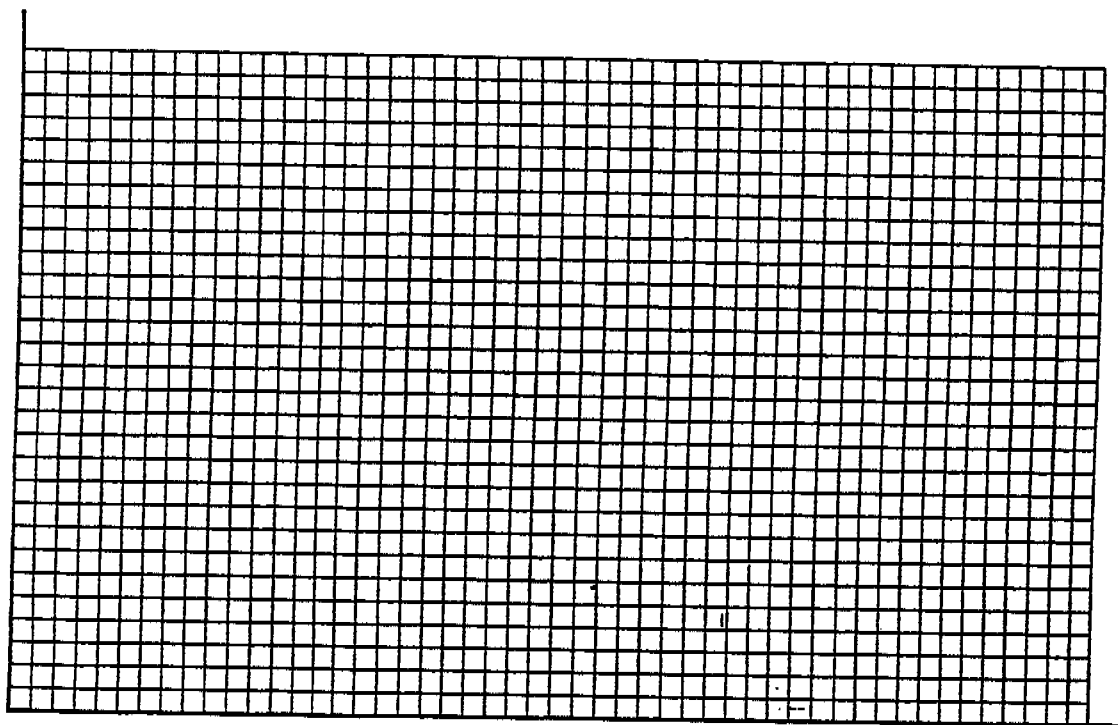
An acid spill occurs during an experiment. A student has the option of putting one of four substances, either vinegar, baking soda, water or salt ($NaCl$), on the spill to neutralize the acid. Which of these materials is the best choice? Explain your answer.

Extended Constructed Response

7. Chemistry

Content CLG: Goal 1, Expectation 5, Indicator 1; Goal 4, Expectation 2, Indicator 4;
Goal 4, Expectation 3, Indicator 1; Goal 4, Expectation 3, Indicator 4
Skills for Success: 3.1

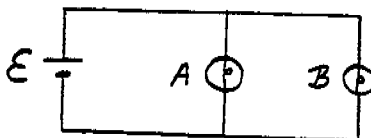
On the axes below, construct a heating curve for water in the form of a line graph. Begin with the solid phase and include all three phases. The axes of the graph should be labeled and should also clearly indicate the temperatures at which phase changes occur.



Describe the changes in the motion of the molecules of water as the temperature increases. Relate these changes in molecular motion to the shape of the heating curve.

Selected Response

Questions 1-2 refer to the following information and set of answer choices.



In the diagram above, two identical light bulbs are powered by an ideal battery having emf \mathcal{E} and no internal resistance.

- A. It becomes brighter.
- B. It stays at the same brightness.
- C. It becomes dimmer but does not completely go out.
- D. It completely goes out.

1. **Physics**
Content CLG: Goal 1, Expectation 2, Indicator 5; Goal 5, Expectation 2, Indicator 2
Skills for Success: 2.2

If bulb B is removed from its socket, what happens to bulb A ? [Key: B]

2. **Physics**
Content CLG: Goal 1, Expectation 2, Indicator 5; Goal 5, Expectation 2, Indicator 2
Skills for Success: 2.2

If bulb B is shorted, what happens to bulb A ? [Key: D]

Brief Constructed Response

3. **Physics**
Content CLG: Goal 1, Expectation 6, Indicator 3; Goal 5, Expectation 3, Indicator 1
Skills for Success:

If 800 grams of water at 40°C are mixed with 400 grams of water at 70°C , what is the final temperature of the mixture, assuming no heat is lost to the surroundings?

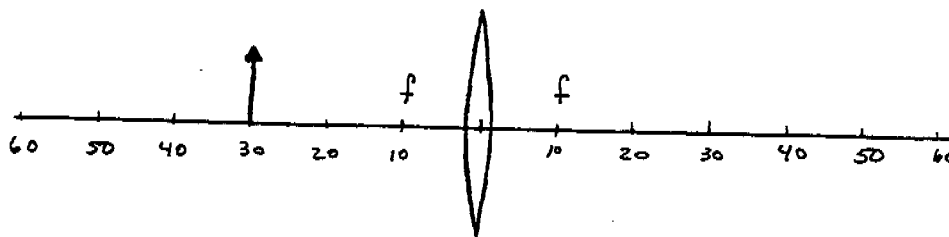
4.

Physics

Content CLG: Goal 1, Expectation 4, Indicator 8; Goal 5, Expectation 4, Indicator 4
Skills for Success: 1.3, 2.2, 2.4, 3.1, 3.2

An object is placed 30 cm from a convex lens of focal length 10 cm.

- a) On the diagram below, draw rays to locate the image that is formed by the lens.



- b) Indicate in writing whether the image is real or virtual.
c) Indicate in writing whether the image is erect or inverted.
d) Using the lens equation which appears on the answersheet, calculate the position of the image.

5.

Physics

Content CLG: Goal 1, Expectation 6, Indicator 3; Goal 5, Expectation 5, Indicator 1
Skills for Success:

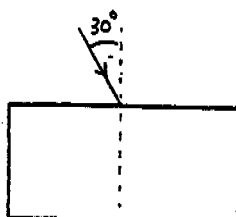
A student measures the radiation emitted by a light source and detects electromagnetic waves of wavelength 7.0×10^{-7} meters. Determine the energy, in joules, of a photon of this light.

6.

Physics

Content CLG: Goal 1, Expectation 6, Indicator 3; Goal 5, Expectation 3, Indicator 1
Skills for Success: 3.1

A ray of light traveling through air is incident upon a sheet of crown glass as shown. The index of refraction for crown glass is 1.52.



- a) On the figure, draw a ray to represent the refracted light in the glass.
b) Calculate the angle of refraction of the light at the surface of the glass.

Extended Constructed Response

7. Physics

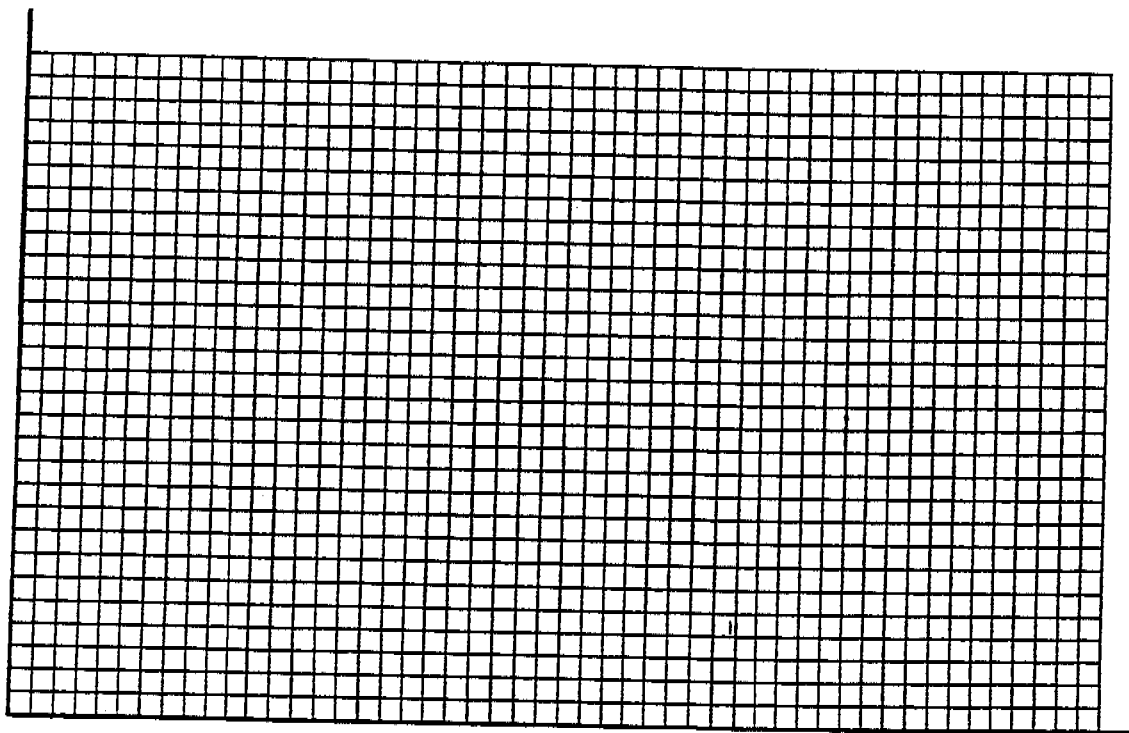
Content CLG: Goal 1, Expectation 2, Indicator 5; Goal 5, Expectation 2, Indicator 1;
Goal 5, Expectation 2, Indicator 2

Skills for Success: 2.2, 3.2

The force, in millinewtons, between two small spheres was measured as a function of their center-to-center distance, and the data obtained are listed in the table below.

Force (mN)	100.0	25.0	11.1	6.3	4
Distance (m)	1	2	3	4	5

- a) On the axes below, plot the data using distance as the independent variable.
Draw a curve that is your estimate of the best fit to the data points.



- b) Based on your graph, describe in words the variation of force with distance.
c) Based on your graph, write an equation that indicates how the force varies with distance. Express this relationship in the form of a proportionality.
d) List two forces in nature which exhibit this mathematical relationship.

SOCIAL STUDIES

Set containing Selected Response and Brief Constructed Response

Questions 1 and 2 refer to the cartoon below.



1. U. S. History

Content CLG: Goal 1, Expectation 2, Indicator 4

Skills for Success: 2.2.1

a) Circle the decade in which you believe this cartoon was drawn. [SR]

1920's

1940's

1960's

1980's

b) Citing specific historical evidence, explain why you chose the decade you did. [BCR]

Key [SR]: 1960's

2. U. S. History

Content CLG: Goal 1, Expectation 2, Indicator 4

Skills for Success: 2.2.1

What is the main message of this cartoon? [BCR]

Selected Response

3. U. S. History
Content CLG: Goal 1, Expectation 1, Indicator 6
Skills for Success:
- What was one goal of the New Deal?
- A. To increase farmland by cutting down trees
 - B. To increase both industrial and farm surpluses
 - C. To increase taxes for farmers and businessmen
 - *D. To increase the American consumer's purchasing power
4. U. S. History
Content CLG: Goal 1, Expectation 2, Indicator 5
Skills for Success:
- Chart 1 (on the next page) shows a change in immigration. Which of the following best explains this change?
- A. Rise of European communism
 - *B. U.S. laws restricted "targeted" immigrants
 - C. Lack of employment opportunities in the U.S.
 - D. Prosperity in Europe after World War I
5. U. S. History
Content CLG: Goal 1, Expectation 2, Indicator 2
Skills for Success:
- Which situation best accounts for the differences in Federal income and spending between 1928 and 1936, as shown in Chart 2 (on the next page)?
- *A. Government funding of programs to combat economic problems
 - B. Increase in personal income tax rates
 - C. Military spending for World War II
 - D. United States trade imbalance with Japan

Chart 1

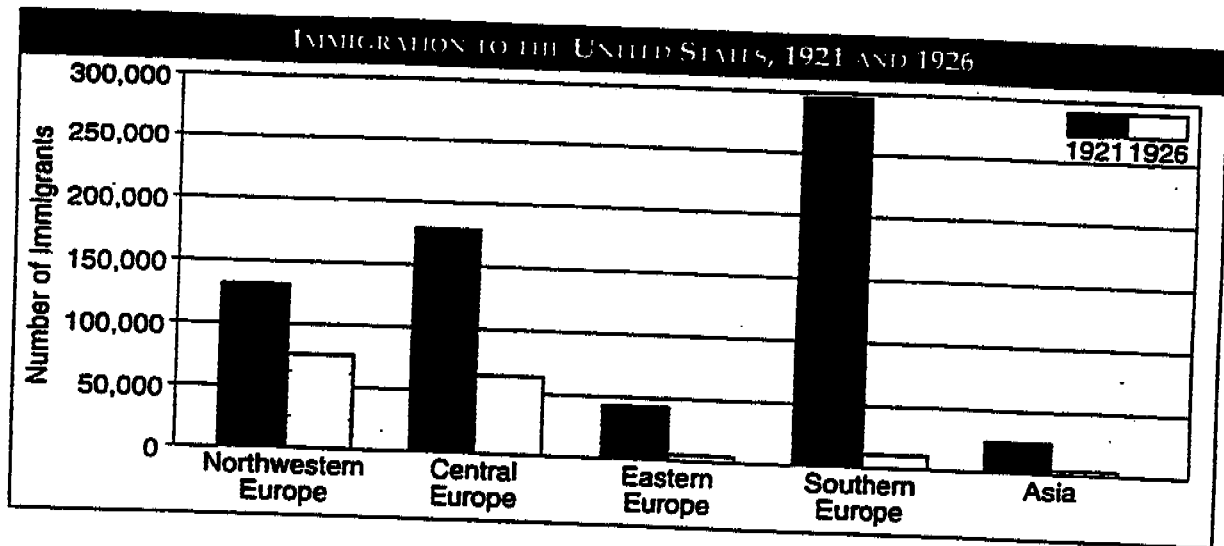
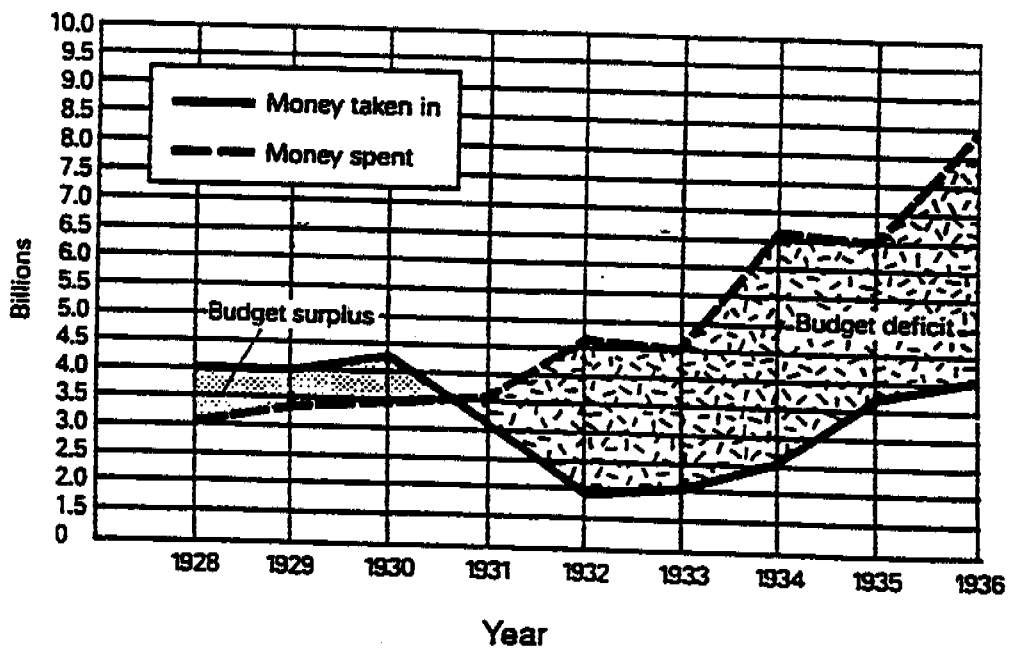


Chart 2

Federal Income and Spending, 1928-1936



6. Government
Content CLG: Goal 1, Expectation 1, Indicator 3
Skills for Success: 4.1

The widespread use of computers has led to a national concern over

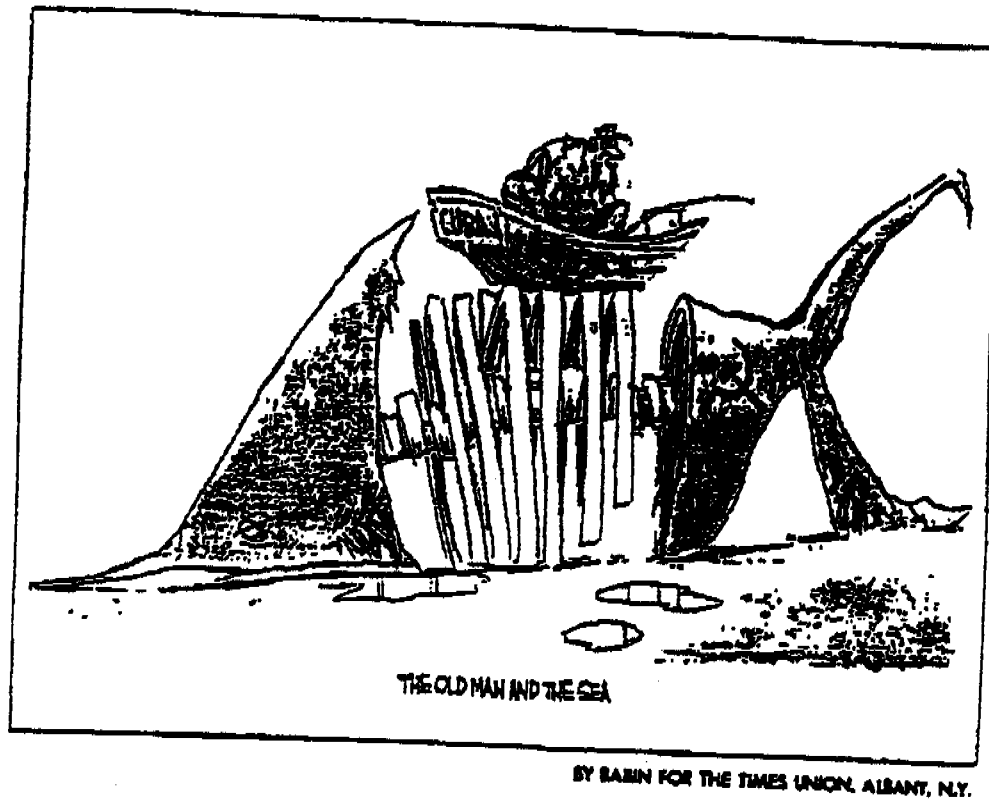
- A. increased pollution of the environment
- *B. guarding the right to privacy
- C. protection of the right to petition
- D. a decline in television viewing

7. Government
Content CLG: Goal 1, Expectation 2, Indicator 4
Skills for Success:

Poll taxes and grandfather clauses were devices used to

- *A. deny African Americans the right to vote
- B. extend suffrage to women and 18-year-old citizens
- C. raise money for political campaigns
- D. prevent immigrants from becoming citizens

8. World History
Content CLG: Goal 2, Expectation 3, Indicator 2
Skills for Success:



What is the main idea of the cartoon above from the late 1980s?

- A. Cuba's fishing industry was suffering a decline.
- *B. Cuba was isolated without Soviet economic support.
- C. Castro rode the wave of world communism to a successful conclusion.
- D. Castro was responsible for the failure of communism in Eastern Europe.

9. World History
Content CLG: Goal 1, Expectation 1, Indicator 3
Skills for Success:

During the Middle Ages, Europeans did not eat potatoes or corn because

- A. their consumption was forbidden by the Catholic Church
- *B. they had not yet been introduced to Europe from the New World
- C. they were believed to be poisonous
- D. they were too expensive to import from China

Brief Constructed Response

10. U. S. History
Content CLG: Goal 4, Expectation 1, Indicator 4
Skills for Success: 2.2.3

What was a weakness in the American economic system in the 1920s that eventually led to the Great Depression of the 1930s? Explain your answer.

11. U. S. History
Content CLG: Goal 1, Expectation 1, Indicator 6
Skills for Success: 2.2.1

Describe a New Deal program that attempted to correct the problem of unemployment during the Great Depression.

12. U. S. History
Content CLG: Goal 3, Expectation 2, Indicator 3
Skills for Success: 2.2.3

How did the geographic location of the United States contribute to the isolationist policies of the 1930s? Explain your answer.

Extended Constructed Response

13. U. S. History
Content CLG: Goal 1, Expectation 1, Indicator 6
Skills for Success: 2.2.6

What effect did the Depression have on families? Use both passages below on two people's memories of the early years of the Depression and your knowledge of history in your answer.

I. Oh, I remember having to move out of our house. My father had brought in a team of horses and wagon. We had always lived in that house, and we couldn't understand why we were moving out. When we got the other house, it was a worse house, a poor house. That must have been around 1934. I was about six years old.

II. The oil boom come [to Oklahoma City] in '29. People come from every direction in there. A coupla years later, they was livin' in everything from pup tents, houses built out of cardboard boxes and old pieces of metal that they'd pick up-anything that they could find to put somethin' together to put a wall around 'em to protect 'em from the public.

Appendix E

Issues and Recommendations from the Program Specifications Committee May 6, 1997

The Program Specifications (PS) Committee reached consensus on the topics listed below.

Timing

1. Each test should be given on one day.
2. All tests should have the same time limit.

Test Issues

1. The selected-response test questions should have four answer choices, only one of which is correct.
2. The committee recommended a higher level of technology be used or tested. They suggested the following technologies: word processor (for essays), electronic dictionaries, Internet (for answering questions), computerized testing. They recognized that these could not all be used during testing, but they felt that questions could test the understanding of the applications of technology.

Retesting

1. Appropriate assistance must be available for students. If a student does not wish to take advantage of the appropriate assistance, parents should be required to sign a parental agreement form.
2. Local school systems should design appropriate assistance programs. Some possible approaches might be summer programs, special courses, tutoring, after school or weekend classes. Appropriate assistance opportunities are not limited to school or county offerings.
3. The PS committee recommends that there be a summer administration date, in addition to the planned dates.

Validity Issues

1. Security of the materials must be maintained at all times. Reviews of test materials should be conducted under secure conditions such as: a) supervised reviews, b) a time limit for reviews, and c) no copying of materials.
2. Non-MSDE reviewers of fairness should include the following groups: special education teachers, linguistic reviewers, classroom teachers, testing experts, university teachers, members of the business community, parents, and central office staff.
3. Reviewers of content should be knowledgeable in the relevant Core Learning Goals.

Score Reporting

1. The appeal policy should be the same for all tests.
2. The appeal policy should be communicated widely.
3. The appeal process should consist of two components: a) an automatic review of any score that is near the decision point, and b) an appeal process for other students, by request.
4. Performance should be evaluated with a proficiency level, not merely a pass/fail basis.
5. Full disclosure of a test will entail additional costs, so detailed score reports could provide information instead.
6. Information should be available to parents before the tests; this information should include sample items and other descriptive information about the content of the tests (i.e., the CLGs), test format, and the way the scores will be used.

Student Populations

1. Out-of-state transfer students will be required to take the test for courses that have been taken in Maryland public schools.
2. Students who transfer from private school to public school will be required to take the test when the courses have been taken in Maryland public schools.
3. Students who transfer between districts in Maryland should, in general be required to take the test. However, there are areas still to be resolved: a) students who transfer in or out of integrated math courses, and b) students who transfer from a full-year course to a semester-length course.
4. There should be no exemption for students taking Advanced Placement courses or students in the International Baccalaureate program.
5. Students should be able to take an HSA test without having ever taken the course, so long as they have approval from a local school official. If students take the test without taking a course, they should not receive course credit.
6. If the HSA testing program is separate from the courses, these recommendations may need to be revisited.
7. Procedures for students with a temporary disability should be the same as used for the MSPAP guidelines.

Special Education Populations

1. The accommodations used for the test must be the same as those used for instruction.
2. The PS committee recommended a meeting of special education directors who should set the procedures for special education students. Among the issues to be resolved by this group would be: a) special education students do not have access to all courses, so is it appropriate that they be required to pass the tests; b) the tests might increase the drop-out rate for special education students; c) there could be staffing problems for actual test administrations because there will not be enough special education

- teachers available; and d) if the English tests are read and reading is considered an essential skill to be demonstrated, should students get an English grade.
3. For visually impaired students, material that can be Brailled should be Brailled. If material can't be Brailled, it should not be scored.
 4. Student records in the school and MSDE should contain an annotation of the accommodations that have been provided. The student's score report should not contain this information.

Limited English Proficiency (LEP)

1. A meeting of LEP directors should be convened to determine procedures for LEP students. Among the issues to be addressed are: a) information for parents should be available in which native languages, b) how will the state consider courses from foreign countries, and c) should LEP teachers determine when LEP students are ready to test or should another mechanism be used (e.g., TOEFL-like test).
2. Scorers for constructed response questions should receive advice to help them score the papers from LEP students. Special scorers may be necessary.

Recommendations about Working Assumptions

1. Test Specifications assumption 7: The percentage of minority representation should represent the percentage of the minority student population in the public schools through the year 2002.
2. Program Specifications assumption 5: There should be three test administrations per year.
3. Program Specifications assumption 8: Accommodations should be available if either an IEP or a 504 plan requires one.
4. Test Administrations assumption 4: The assumption should state that tests are given within the normal school week since summer schools do not always have Friday classes.

Issues Raised By Program Specifications Committee

Timing

- About half of the committee felt that at least one ten-minute break should be given; about half felt that there should not be a scheduled break. Any break time would not be considered part of the testing time.

Tests

- Most participants felt that the tests should be given on one day because there are serious operational problems in administration over many days (e.g., absences, matching of materials for scoring, etc.). A minority felt that two days would be less fatiguing for students.
- Many participants were concerned about the balance of multiple-choice and constructed-response questions. They wanted more constructed response.
- Some participants felt that there should be no assumption that the tests will be used as a graduation requirement.

Pretesting

- There are advantages and disadvantages to pretesting within the operational test. Many participants wanted to avoid including any additional multiple-choice questions; however, some participants felt that the added number would be low and would help strengthen the test pool without inconvenience.

Retests

- Many participants felt that students should be required to take remediation.
- Many participants felt that students should be required to take the test when they take the course.
- Since participants were unsure whether the test would count toward part of the course grades, they felt unable to give information on some topics.

Validity Issues

- Some participants felt that reviewers outside MSDE should receive staff development training before they do reviews.
- Some reviewers felt that it was not necessary to have content reviews by non-MSDE staff.
- Some participants felt that groups represented in non-MSDE reviews should have input to the cut score decision.

Score Reporting

- Some participants felt that there should be disclosure of a complete test.
- Some participants recommended that only Pass/Fail scores be given because special merit can be assessed in other ways. Others felt that they wanted proficiency level information.

Student Populations

- Participants wanted to know whether the test will be used as part of course credit.
- Some participants felt that out-of-state students should have the opportunity to take the tests if “merit” is linked to the tests.
- Some participants felt that the test should be required of home- and hospital-schooled students if they are seeking a MD diploma.
- Some participants felt that transfer students who transfer when less than half of the course has been completed should not have to take the test; but if more than half the course has been completed, they should take the test. Others felt that the school principals should have the discretion to exempt any transfer student; in this case, guidelines should be provided.

Limited English Proficient (LEP)

- Some participants felt that those in “sheltered English” classes shouldn’t be held accountable for the test.
- Some participants felt that local alternatives to the test should be available.
- One participant suggested the following possible accommodations: a) students should be able to ask the test administrator for synonyms for words outside the content area. b) words not in a bilingual dictionary that are used on the test should be glossed, c) students should be able to bring their own word book of new words with them to the test. Some participants disagreed with each of these suggestions because they would provide LEP students with advantages not held by other test-takers.
- One participant asked whether timing would be adjusted for LEP students if accommodations, such as bilingual dictionaries, are available.
- One participant asked whether any action had been taken with regard to a working paper on testing of special populations that had been submitted to MSDE.

Other

- There is concern on the part of some PS members that there are lots of tests required in schools at the end of semester and that the HSA tests pose an additional time burden.
- Some PS members felt that an upper limit should be placed on the graphing calculator functions in the Math tests.
- A question was asked: do review committees have the power to delay operational administrations if they are dissatisfied with the test questions?
- Some committee members felt that schools cannot currently provide access to the technology that is being recommended (e.g., word processors, computers) so they felt there would be unfairness if the tests relied on this equipment.

Issues To Be Resolved

- Is it the case that only some of the Skills for Success can be tested (e.g., thinking/communicating)? Or can test questions serve as surrogates for some of the Skills?
- Will there be formula scoring of tests (i.e., a penalty subtracted from the number correct for wrong answers)?
- How many no-fault administrations will be given?
- Can students delay testing (e.g., take the Math 1 test after Math 2)?
- Can students take tests in 8th grade if courses are taken then?
- How should the use of selected-response pretest items be handled on operational tests? Can some items be included in the operational tests or must all items be pretested outside the test?
- Can students take a test without having taken a course? The answer to this question is dependent on whether the test counts as part of a course grade. Can anyone help determine whether tests will count toward course grades?
- What kind of scale should be used for score reports (i.e., Pass/Fail or proficiency levels)?
- How should home-schooled students be treated?
- Should there be alternatives to the HSA tests that could certify knowledge of the CLGs? If so, what should they be? If so, should students be expected to meet the same standard as for the HSA tests (and how can that standard be established) or should students find the alternative path harder so that they are discouraged from taking the alternative path?
- Special Education Populations: Does the PS committee wish to make any recommendations with regard to the issues it raised at the last meeting?
- Limited English Proficient Populations: Does the PS committee wish to make any recommendations with regard to the issues it raised at the last meeting?

Issues and Recommendations from the Test Administration
Specifications Committee¹
May 6, 1997

The Test Administration Specifications (TAS) committee had two meetings and reviewed a number of general issues concerning the design of the 12 assessments and the operational implementation of the High School Assessment program. Issues presented to the Test Administration Specifications Committee were identified by the MSDE and contractors collaboratively and provided to the committees for discussion at the meetings. Committee members reviewed and discussed most of these issues; they also tabled some issues that they believed could be addressed later in the test development process and identified additional issues for discussion at the meetings. Because these meetings were not sufficient to adequately address all issues which will impact the design and implementation of the high school assessments, additional meetings of this group or subgroups will be required during the next six months.

Members of the Committee included: George Newberry (Co-Chair), Gary Dunkleberger (Co-Chair), Wayne Camara (College Board staff and recorder at the second meeting), John Fremer (ETS staff and recorder at the first meeting), Mary Beth Adams, Vicki Carter, Esther Collier, Mary Douds, Maureen Beaupre, Donna Faith, Ruth Ann Hall, Roberta Hays, Jane Higdon, Bruce Hislop, Julian Katz, Karen Kunkel, Paul Mazza, Ruth Orland, Donald O'Neal, Robert Pfau, Debra Slider, Jo Sowers, Leroy Thompkins, Edward Weiland, Leslie Wilson, and James Younkens.

At the beginning of the second meeting Wayne Camara attempted to contrast two responsibilities the College Board and ETS have regarding the development of recommendations concerning the test administration specifications that will be submitted to the Maryland State Board of Education. He noted the first responsibility of the contractors is to serve as the recorder, documenting all recommendations that are reached through a consensus process, while the co-chairs facilitate and direct the meeting. The second responsibility the contractors have is to the MSDE and MSBE to elaborate on the administrative recommendations forwarded by the committee, both in terms of how they might be operationally implemented, but also to identify major weaknesses or risks in any recommendations that could threaten the validity of the assessments, increase operational costs substantially or introduce serious threats to test security or standardization which could threaten the fairness of the assessments. He noted that there administrative features should balance the educational and instructional rationale with the psychometric, legal, and financial constraints for a high stakes individual accountability assessment program. Recommended features which can meet both sets of priorities, even when they

¹ This summary was completed by Wayne Camara, The College Board, and reflects notes and summaries from the meeting on issues where there appeared to be a consensus. Individual members of the test administration committee and staff from the College Board and Educational Testing Service may not necessarily agree with all recommendations or even agree that they represent all items of consensus that were reached by the committee.

impose additional costs and burdens to the state, should be identified separately from recommended features which may not adequately meet one set of these priorities.

Mr. Camara attempted to briefly comment on any potential recommendations which might not address the psychometric constraints or increase costs to the program during the discussion so members of the committee would be aware of all consequences of potential administrative features. These comments were not meant to deter the committee from pursuing its own independent recommendations, but to inform them of areas where the College Board and ETS may need to propose alternative recommendations to the Board in order to fulfill their contractual obligations as technical and psychometric design consultants. Below, an '*' is used to indicate where significant concerns were raised with the TAS recommendation, or a specific portion of the recommendation, by the College Board staff. These concerns will be elaborated in the final report. A '**' indicates that more specific guidelines are required from TAS to complete the HSA design and describe the implementation of the recommendation. In most of these instances the TAS either had insufficient time to develop more specific recommendations or chose to defer the work until after the design work has been completed.

In general, the committee felt strongly that each school should have maximum flexibility in all aspects of preparing for and administering the High School Assessment Program to the extent possible. Where such flexibility or lack of standardization might compromise critical issues such as test validity and test security, or substantially increase costs, MSDE have a responsibility for conducting additional studies and investigating a variety of options for implementing the committee's recommendations before reaching conclusions about their psychometric, legal or financial consequences. The committee also felt that a long-term goal should be to move to computer-based testing, and that additional consultation is needed to determine how to transition many of the constructed response items required by the current design to this administrative platform in the future. The committee felt that computer-based testing could provide additional ways of dealing with many of the issues raised by the committee, but realized that item types preferred by the separate test specifications committees have not been implemented on this medium to date.

A number of specific recommendations and issues were discussed which can be classified into the following categories: (1) General Issues, (2) Student Populations, (3) Equipment, (4) Administration Timing, (5) Scheduling, (6) Security, (7) Preparation Plus Design, and (8) Other.

(1) General Issues

1. No student ID should be required for students enrolled in public high schools completing the assessments, but each school would have responsibility to develop and implement procedures which verify the actual student is taking each test. School principals and superintendents would be held accountable for developing

their own procedures which ensure the appropriate student is completing the appropriate test. **

2. Public schools should only be required to administer assessments for students enrolled in their schools. Alternative administrative mechanisms should be developed for home-bound schools or private school students if and when they are to participate in the assessment program.
3. Students should be tested in groups that schools see as appropriate. Administration conditions should permit schools maximum flexibility to administer assessments within individual classes (with in tact students or other student assignments to specific classrooms), large-scale test administration (e.g., study hall, cafeteria), or other settings as long as they adhere to or exceed state guidelines. *
4. Staff development must be funded and occur prior to the first no-fault administration of the assessments to ensure teachers are prepared to provide appropriate required instruction in the Core Learning Goals and that students have had an opportunity to learn (e.g., teachers need training on appropriate graphing calculators well before the math assessments are administered since these calculators will be required for the test).
5. Defer development of an administrative manual for the high school assessments until this summer when a subcommittee of the TAS will be assembled. The manual should be modeled after the existing MSPAP guidelines and procedures for administrators and proctors. **

(2) Student Populations

1. Students should be allowed to take two tests from the same content area (e.g., English 1, English 2; Biology, Physics) during the same test administration (i.e., fall, spring).
2. Defer decisions on accommodations until a subcommittee of TAS can be assembled to revise the 'Requirements and Guidelines for Exemptions, Excuses and Accommodations for Maryland Statewide Assessment Programs' document. **

(3) Equipment

1. The state will provide the necessary funds to aid local districts in purchasing the number of graphing calculators required for the high school assessment before the beginning of the school year in which the first no-fault administration will occur.

2. MSDE should commission a study during the field trials which will examine the effects of different types of graphing calculators and students' experience in using graphing calculators. Results must demonstrate that there are no significant differences in student performance due to differences in the calculator used or student exposure and familiarity with calculators prior to issuing individual student scores on the math assessments.
3. There will be discrepancies between the existing equipment in schools and what will be required for instruction and assessment. A subcommittee of the TAS should be formed this summer to develop a proposal to support a request from the General Assembly for the necessary equipment required for the high school assessment. The TAS was unable to identify additional equipment needs because the test specifications committees have not fully come to closure on the requirements for equipment.

(4) Administration Time

1. All students must have the maximum amount of time available to complete the tests. Tests can not be speeded. Results from field testing must be reviewed by the TAS and test specifications may need to be required when results are available.
2. If possible, tests should be untimed to permit all students to complete the entire test at their own pace as is currently provided with the Functional Testing Program. *
3. Each test will include 165 minutes of testing time, 15 minutes of instructions and up to two breaks.
4. The duration of the breaks, the number of breaks (1 or 2) and the point in testing time where the break occurs must be left to each individual school to determine. Some members felt that individual teachers should determine the time of the break. *

(5) Scheduling Issues

1. Schools will determine whether or not they will close/cancel classes on HSA testing dates for students completing the assessments and or all students. **
2. All students across the state will be expected to take the same assessments (e.g., English 1, Math 2) on the same date(s), but each school can determine the administration times and other parameters of the local schedule.
3. Two administrations (winter and end of year) for each test is reasonable, but the MSBE must be prepared to justify this to districts which have half-credit courses

or 1/4 courses since these students will be penalized by the limited number of administrations.

4. Only certified or provisionally certified teachers should be allowed to serve as proctors. Some schools may designate the classroom teacher as proctor for some or all of their classes, while other schools may use substitute teachers of other subject teachers for this function.**
5. The LAC's will determine appropriate procedures for training proctors at their schools. **
6. There is no consensus among TAS that 12 tests Prep Plus can be successfully administered in schools twice each year without dramatic interruptions in the school calendar and schedule and a few prototype schedules should be developed by MSDE to illustrate potential schedules before proceeding further.

(6) Security

1. A pamphlet should be produced by each district, based on state guidelines and policies, that would outline appropriate and inappropriate student conduct in preparation for and completion of the tests. This brochure would be signed by each student and provide documentation that they have been informed of the policies and consequences of inappropriate test taking behavior. Parents should receive a copy of the brochure as well.*
2. Additional materials describing the HAS must be immediately developed and shared with parents, students, teachers and the general public. These materials should be distributed to students by middle school or earlier and at the beginning of each year that assessments may be required.
3. Spiraling (scrambling) of items is desirable if it will increase test security and not interfere with test performance in any way. Content teams should review all proposed spiraling to ensure that it will not interfere in the nature of the tasks.
4. Defer specific policies concerning security, test site conditions, and analysis for cheating until MSDE staff review existing policies for MSPAP and provide the TAS with proposed changes and modifications. **

(7) Preparation Plus (English exams only)

1. All English assessments should be administered and completed on one-day and not require two days for administration as proposed by the test specifications committee.

2. Preparation (of 60 minutes) would ideally occur the day of the assessment, but may occur a few days prior to the assessment if required for school logistical operations.*
3. There is no consensus among TAS that tests requiring Prep Plus can be successfully administered in schools and a few prototype schedules should be developed by MSDE to illustrate potential schedules before proceeding further.
4. Students who miss preparation activities or a portion of the assessment must be allowed to make-up the preparation or assessment during the make-up days of the administration. **
5. Further investigation is required to determine whether adequate numbers of certified teachers will be available to conduct the preparation activities. Schools should be permitted to determine how preparation is provided and by whom (e.g., in-tact classes by teacher or other teacher, large-groups).
6. Defer all other operational issues regarding Prep Plus administration to the English test specifications committee for consideration. **

(7) Other (Cost Implications, etc.)

1. State and local support is required for the purchase of all equipment, staff development, and support for the administration of assessments. This support is required prior to the commencement of school in the year of the first state-wide no fault administration. If local districts are to purchase equipment they will need funding and detailing information one year in advance of this (or by January, 1998).
2. A minimum of one dedicated staff person will be required at each school and it is expected that local districts will be responsible for these costs.
3. A number of special studies must be designed (e.g., speededness, varying number of MC items which can result in a reliable test, calculator differences) incorporated into the upcoming no-fault administrations and any pilot testing that is conducted. All studies must include representative groups of students from special needs and LOEP populations, and include geographically diverse schools. **
4. Local districts will be required to pay for additional substitute teachers needed for successful administration of prep plus and the assessments and detailed information on the administrative requirements in needed to help locals prepare for these requirements. **

5. Local districts will also be responsible for developing remedial classes and other appropriate programs for student who do not pass the required tests. *
6. The state will be required to provided financial support for the development of all tests, printed materials describing the tests, scoring and score reporting, special research studies, and a prototype test that can be shared with educators in the next year. State will assist locals in supporting remediation, equipment purchases, and staff development expenses. Local districts will support staff and substitute teachers required for administration and instruction, and additional staff required to manage the school's assessment scheduling and administrations.

TEST SPECIFICATIONS COMMITTEE MEMBERS

ENGLISH

Co-Chairs:

Mary Jo Comer
Trudy Collier

English I

Conrad, Deborah
Edwards, Sylvia
Rice, Barbara
Schultz, Gretchen
Shaw, Patricia
Veslany, Lisa
*Haspel, Paul

St. Mary's
Anne Arundel
Washington
Queen Anne's
Montgomery
Howard
Cecil Community College

English II

Cathell, Joanne
Gelsinger, Barry
Jenkins, Kathleen
McGraw, Hank
Nored, Deanna
McNair, Joann
*Gross, Monica

Worcester
Carroll
Charles
Baltimore County
St. Mary's
Baltimore City
Bowie State University

English III

Bond, Linda
Felix, Georgia
Finan, Frank
Hudson, Elise
Smith, Karen
Tumminello, Anelle
*Glaser, Michael

Allegany
Baltimore City
Calvert
Cecil
Somerset
Anne Arundel
St. Mary's College

Scorers:

David Nuzzi
Shirley Phillips
Sheila Vaughn

Cecil
Prince George's
Prince George's

*Indicates representative from Higher Education.

TEST SPECIFICATIONS COMMITTEE MEMBER

MATHEMATICS

Co-Chairs: Elaine Crawford
Cindy Hannon
Leslie Hobbs (Washington Co.)

Algebra

Angel, Sherryl	Worcester
Halstead, Robert	Calvert
Michael, Kevin	Calvert
Mills, Paul	Baltimore City
Waite-Jaques, Barbara	Montgomery
Ward, Bonnie	Frederick
Wise, Roberta	Charles
Booth, Penny	Baltimore County
* <u>Whitehead, Gladys</u>	Prince George's Community College

Geometry

Brown, Martha	Prince George's
Cepaitis, Theresa	Anne Arundel
Deem, Elizabeth	Garrett
Dreschler, Terry	Somerset
Kaniecki, Linda	Harford
Walsh, Charles	St. Mary's
* <u>Graeber, Anna</u>	University of Maryland, College Park

Scorers:

Carol Sander	Montgomery
Maureen Stockman	Caroline

* Indicates representative from Higher Education.

TEST SPECIFICATIONS COMMITTEE MEMBERS

SCIENCE

Co-chairs: Gary Hedges/Diane Householder
Linda Musial (Charles County)

Biology

Bundy, Karen	Allegany
Hohnke, Larkin	Frederick
Thompson, Domenic	Baltimore City
Thornton, Michele	Frederick
Wilson, Beverly	Worcester
* <u>Philippides, Judith</u>	Towson State University

Chemistry

Hillsman, Doria	Montgomery
Kistler, Suzanne	Calvert
Kroeger, Vickie	Montgomery
Lonie, Richard	Cecil
Newsome, Demetria	Baltimore City
* <u>Harwood, William</u>	University of Maryland, College Park

Earth/Space Science

Andrione, Ruth	Baltimore County
Barnes, Ronald	Baltimore County
Green, Margie	Prince George's
Lehman, Brian	Harford
Mowrer, Nancy	Talbot
Luniewski, Karen	Carroll
Szesze, Michael	Calvert
* <u>Henry, Richard</u>	Johns Hopkins University

Physics

Cox, Brenda	Wicomico
Hopkins, Barry	Anne Arundel
Mayfield, David	Garrett
Visintainer, Carol	Caroline
* <u>Treacy, Donald</u>	Naval Academy

Scorers:

Lynn Jones	Carroll
Meredith Quinn	St. Mary's
Daniel Richardson	Worcester
Leslie Rogers	Montgomery

* Indicates representative from Higher Education.

TEST SPECIFICATIONS COMMITTEE MEMBERS

SOCIAL STUDIES

Co-Chairs: Diane Johnson
Barbara Graves (Charles Co.)

United States History

Bunitsky, Michael	Frederick
Gift, Edward	Washington
Sowders, William	Howard
Thompson, Ronald	Cecil
Vandenburg, Paul	Anne Arundel
Wehrle, John	Montgomery
* <u>Christian, Charles</u>	University of Maryland, College Park

Government

Altoff, Peggy	Carroll
Carr, Clementine	Baltimore City
Ellis, Mavis	Montgomery
Oliver, Kaye	Calvert
Mister, Coleen	Worcester
Roche, Kay	Allegany
* <u>Jenne, Joel</u>	Salisbury State University

World History

Jones, Mary Louise	Allegany
Marshall, Margaret	Charles
Prewitt, Joann	Baltimore City
Shepard, Rex	Baltimore County
Shepherd, Anita	Calvert
Springle, Mary	Worcester
* <u>Yip, Ka-Che</u>	University of Maryland, Balto County

Scorers:

John Abel	Cecil
Marjorie Jenkins	Montgomery
Alex Spooner	Harford

* Indicates representative from Higher Education.

The following is the make-up of the Program Specifications Committee:

English content:

Linda Flannigan	4 participants
Allan Starkey	Charles
Kathryn Harcum	Howard
Pamela Enrico	Caroline
	Calvert

Mathematics content:

Frances Albert	3 participants
Janice Williams	Howard
Sharon Wiggs	Montgomery
	Baltimore City

Science content:

Claudia Wortman	3 participants
Brad Yohe	St. Mary's
James Strandquist	Carroll
	Prince George's

Social Studies content:

Crawford Clark	3 participants
Patsy Somers	Montgomery
Charles Ridgell	Somerset
	St. Mary's

Participants with expertise in scoring, particularly in large scale assessment and the unique requirements of the scoring process:

Sandra Baker	4 participants
Donald DeMember	Caroline
Caroline Hall	Montgomery
Kara Libby	Worcester
	Prince George's

Gifted and Talented (appointed by Deborah Bellflower):

Diane Sprague	3 participants
Donna Chesno	Anne Arundel
William Patton	Allegany
	Queen Anne's

Special Education (appointed by Carol Ann Baglin):

Kim Williams	3 participants
Frances Collins	Harford
Karen Salmon	Prince George's
	Talbot

"Skills for Success" (appointed by Katharine Oliver):

Linda Cunningham	4 participants
Marilyn Glass	AAI Corporation
Yvonne Moten	Allied Signal Technical Services Corp.
Harry Olson	Baltimore Gas and Electric Co.
	Morris and Ward International, Inc.

LEP (appointed by Jill Bayse):

3 participants

Eunju Chung
Jody Crandall
Marjorie Rosenberg

Baltimore City
UMBC (Director of ESOL Program)
Montgomery

Assistant Superintendents (appointed by Dr. Trader):

3 participants

Mary Helen Smith
John O'Connell
Clarissa Evans

Montgomery
Calvert
Baltimore City (CO-CHAIR)

Superintendent (to be appointed by Terry Greenwood of PSSAM):

1 participant

Spicer Bell

Dorchester

Middle School Personnel (appointed/nominated by Karl K. Pence of MST A):

3 participants

Mary Beth White
Betty Drew
Bonnie Barnes

Anne Arundel
Kent
Wicomico

The following is the make-up of the Test Administration Specifications Committee:

Principals/Assistant Principals:

Karen Kunkel
Robert Pfau
Edward Weiland

3 participants
Charles
Harford
St. Mary's

Guidance Counselors:

Mary Douds
Debra Slider
Vicki Carter

3 participants
Garrett
Allegany
Somerset

Teacher/Test Administrators:

Bruce Hislop
Donald O'Neal
Roberta Hays

3 participants
Charles
St. Mary's
Harford

System Testing Coordinators:

Jane Higdon
Esther Collier
Mary Beth Adams

3 participants
Charles
Worcester
Dorchester

LAC's (elected by ballot from within their own group):

Leslie Wilson
Julian Katz
Paul Mazza

3 participants
Howard
Frederick
Baltimore County

Data Processors (nominated by CSPEIS):

Jo Sowers
Maureen Beaupre
Ruth Orland

3 participants
Washington
Anne Arundel
Montgomery

Assistant Superintendents (appointed by Dr. Trader):

Gary Dunkleberger
Leroy Thompkins

2 participants
Carroll
Prince George's

Middle School Principal (appointed/nominated by Karl K. Pence of MST A):

Donna Faith

1 participant
Frederick

High School Teachers (appointed/nominated by Karl K. Pence of MST A):

James Younkings
Ruth Ann Hall

2 participants
St. Mary's
Charles

Appendix F

SAMPLE BOOKLET INFORMATION

Possible Policy on Testing Aids

Students may find a watch to be helpful, but they may not use a watch with an audible alarm.

Students may not take any of the following into the testing room:

- food or drink
- scratch paper
- notes, books, dictionaries
- highlighters or any kind of colored pens or pencils
- portable listening devices (with or without earphones)
- portable recording devices (with or without earphones)
- photographic equipment
- equipment not supplied by MSDE (NOTE: Math is requiring the students to bring their own graphing calculators.)

Possible Test Security Policy

The following procedures or regulations will be enforced so that all students have an equal opportunity to do their best on the test. Students may be dismissed from the testing room if they fail to follow any of the testing instructions given by the test administrator, or if they:

- attempt to take the test for someone else
- give or receive assistance during the test
- attempt to remove from the test room any part of a test book, answer sheet, or notes relating to the test (either on paper or on a calculator, or in any other electronic form)
- create a disturbance for other test-takers
- look through the test before the start of the test
- read or work on one section during the time allowed for another section
- continue to work after time is called
- leave the test room without permission
- use any of the prohibited aids
- take food or drink into the test room

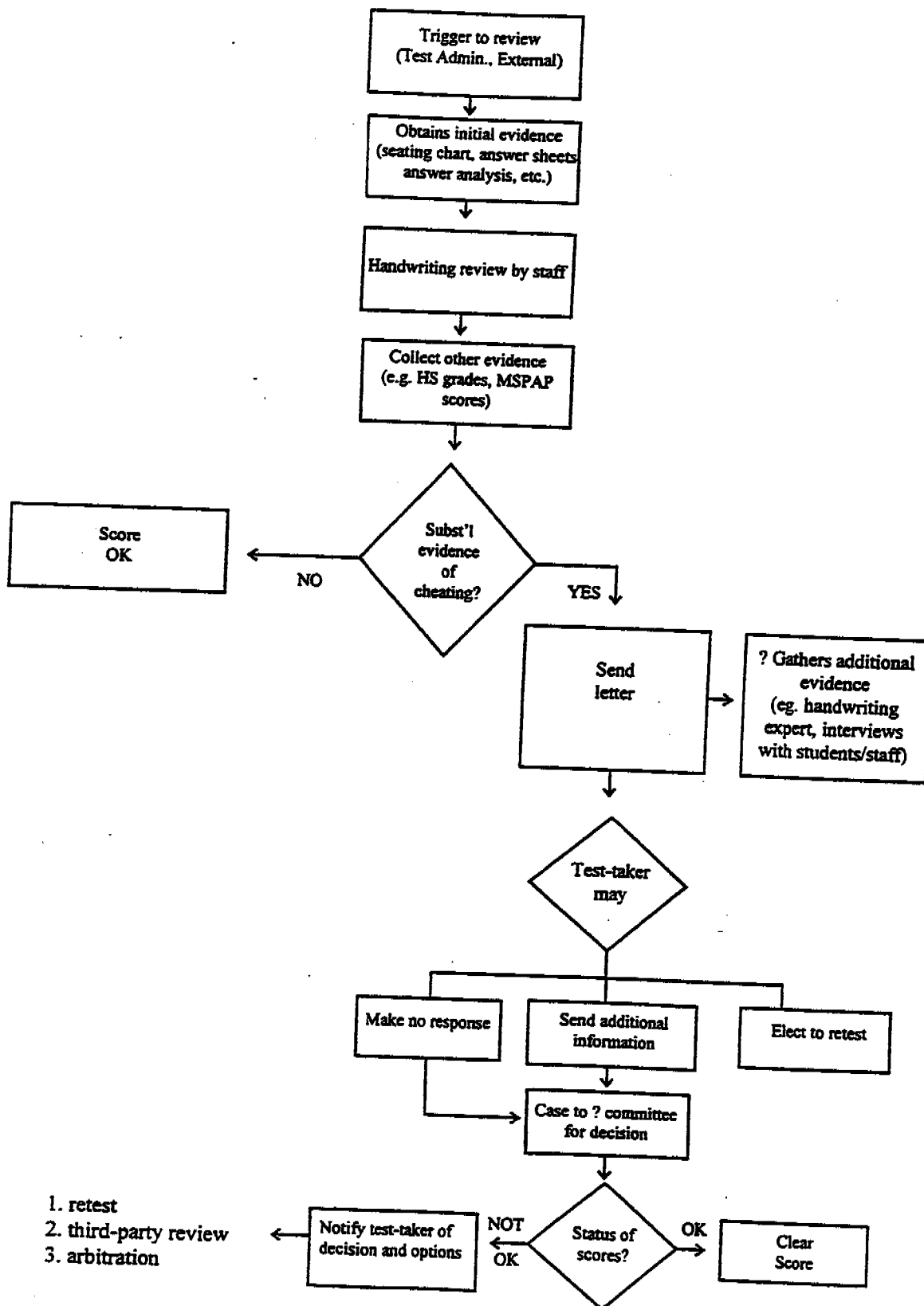
Possible Policy on Questions of Score Validity

Test administration and test security regulations are intended to ensure that all test takers have the same opportunity to demonstrate their abilities and to prevent any student from gaining an unfair advantage over others because of testing irregularities or misconduct. MSDE routinely reviews irregularities and test scores believed to be earned under unusual or questionable circumstances.

MSDE reserves the right to cancel any test score if the student's identification is not valid, if the student engages in misconduct, if there is a testing irregularity, or if there is a good reason to question the score's validity. When test scores are canceled because of group irregularities such as defective materials, improper timing, or preknowledge of the content, students will be given an opportunity to retake the test as soon as possible.

Before test scores are canceled for misconduct, the student is notified and given an opportunity to provide additional information. When the validity of a test score is questioned, the student is notified of the reasons that the score is being questioned and the student has an opportunity to provide additional information. The test-taker may also request a third-party review of the matter.

Processing of Score Validity Inquiries



1. retest
2. third-party review
3. arbitration

**Effects of Scorer Reliability:
Examples from College Board Testing Programs**

In this appendix we illustrate some examples from existing programs of reliability levels and other statistics that have been attained with certain item numbers and combinations of item types. Most of the results are from the technical manual of the College Board's Advanced Placement Testing Program (College Board, 1988).

Effects of Scoring on Test Reliability: College Board Advanced Placement (AP) Tests

The following table presents scorer reliability coefficients that are reported in the AP technical manual. These coefficients are estimates of test score reliability of tests comprised of a mix of SR and CR items. Separate estimates are reported for single reading of each CR item and for double readings.

Table A1. Effects of Scorers on Test Reliability

Subject Area	Year	No. Items SR/CR	Type of Reading	Reliability Estimate
English Lit. & Composition	1984	58/3	single	.84
			double	.91
	1982	55/3	single	.84
			double	.89
English Lang. & Composition	1984	55/3	single	.82
			double	.86
	1982	60/3	single	.84
			double	.90
American History	1981	84/1	single	.79
European History	1981	90/1	double	.90
			single	.78
Chemistry	1980	80/6	double	.88
			single	.95
Physics B	1981	70/6	double	.97
			single	.96
Physics C, Mechanics	1981	35/3	double	.97
			single	.94
Physics C, Elect. & Magnetism	1981	35/3	double	.94
			single	.95
			double	.95

**Test Reliability: College Board AP Tests
Mixes of Item Types**

This section contains reliability estimates for various mixes of item types on certain AP tests. Data were reported by maximum possible score and percentage in the composite rather than number of items of each type.

Table A2. AP Reliability estimates by Item Types

Test	Year	Max Possible Score SR/CR/Comp. (%CR in Comp.)	SR Reliability	CR Reliability	Composite Reliability
English Language and Composition	1982	60/100/150 (67)	.85	.57-.78	.78-.87
	1983	55/90/150 (60)	.76	.62-.75	.78-.82
	1984	65/90/150 (60)	.85	.58-.82	.82-.89
	1985	65/90/150 (60)	.88	.56-.82	.82-.90
	1986	60/90/150 (60)	.84	.60-.82	.80-.88
English Literature and Composition	1982	55/100/150 (67)	.86	.57-.75	.79-.86
	1983	58/90/150 (60)	.81	.61-.75	.80-.85
	1984	60/90/150 (60)	.82	.55-.75	.79-.85
	1985	65/90/150 (60)	.86	.57-.75	.80-.87
	1986	60/90/150 (60)	.86	.56-.75	.81-.87

Table A2 (continued). AP Reliability estimates by Item Types

Test	Year	Max Possible Score SR/CR/Comp. (%CR in Comp.)	SR Reliability	CR Reliability	Composite Reliability
American History	1982	100/30/180 (50)	.89	.60-.79	.84-.90
	1983	100/30/180 (50)	.89	.52-.79	.83-.90
	1984	100/30/180 (50)	.90	.54-.79	.84-.90
	1985	100/30/180 (50)	.89	.54-.79	.83-.90
	1986	100/30/180 (50)	.90	.49-.79	.83-.90
European History	1982	90/30/135 (67)	.90	.46-.63	.71-.79
	1983	100/30/180 (50)	.89	.42-.63	.79-.85
	1984	100/30/180 (50)	.90	.46-.63	.80-.85
	1985	100/30/180 (50)	.91	.44-.63	.83-.87
	1986	99/30/180 (50)	.90	.48-.63	.81-.86
Biology	1982	120/45/150 (50)	.93	.60-.85	.87-.94
	1983	120/45/150 (50)	.93	.68-.85	.89-.94
	1984	120/45/150 (50)	.93	.70-.85	.90-.95
	1985	120/45/150 (50)	.93	.66-.85	.88-.94
	1986	120/45/150 (50)	.93	.73-.85	.89-.93
Chemistry	1982	79/88/160 (55)	.90	.77-.95	.90-.96
	1983	84/88/160 (55)	.91	≥.724	≥.882
	1984	85/88/160 (55)	.90	≥.756	≥.889
	1985	79/88/160 (55)	.92	≥.790	≥.909
	1986	80/88/160 (55)	.91	≥.776	≥.905

Table A2 (continued). AP Reliability estimates by Item Types

Test	Year	Max Possible Score SR/CR/Comp. (%CR in Comp.)	SR Reliability	CR Reliability	Composite Reliability
Mathematics Calculus AB	1982	45/63/126 (50)	.86	.80-.87	.91-.93
	1983	45/45/90 (50)	.90	≥.793	≥.916
	1984	45/45/90 (50)	.89	≥.796	≥.914
	1985	45/50/108 (50)	.89	≥.843	≥.927
	1986	45/54/108 (50)	.90	≥.848	≥.931
Mathematics Calculus BC	1982	45/63/210 (50)	.85	.79-.85	.90-.92
	1983	45/45/126 (50)	.87	≥.735	≥.886
	1984	45/45/90 (50)	.87	≥.711	≥.890
	1985	45/54/90 (50)	.88	≥.753	≥.897
	1986	45/54/108 (50)	.88	≥.804	≥.909
Physics B	1982	70/105/210 (50)	.91	.79-.95	.92-.96
	1983	70/90/180 (50)	.88	.80-.98	.91-.96
	1984	70/90/180 (50)	.90	.85-.96	.93-.97
	1985	69/90/180 (50)	.89	.86-.98	.93-.97
	1986	70/90/180 (50)	.90	.84-.98	.93-.97
Physics C Mechanics	1982	35/45/90 (50)	.85	.80-.96	.90-.94
	1983	35/45/90 (50)	.85	.71-.97	.87-.94
	1984	35/45/90 (50)	.85	.79-.97	.89-.95
	1985	35/45/90 (50)	.84	.70-.97	.87-.95
	1986	35/45/90 (50)	.87	.70-.97	.88-.95

Table A2 (continued). AP Reliability estimates by Item Types

Test	Year	Max Possible Score SR/CR/Comp. (%CR in Comp.)	SR Reliability	CR Reliability	Composite Reliability
Physics C Electricity & Magnetism	1982	35/45/90 (50)	.88	.80-.95	.91-.95
	1983	35/45/90 (50)	.85	.77-.98	.89-.95
	1984	35/45/90 (50)	.86	.67-.98	.86-.96
	1985	35/45/90 (50)	.83	.74-.98	.88-.95
	1986	35/45/90 (50)	.86	.74-.98	.89-.95

Why and How

Educational
Testing
Service

Questions
Test
Scores

Test Security Office
Educational Testing Service
Princeton, NJ 08541
1-800-750-6991
Fax: 1-609-406-9709

Preface

This booklet explains why and how ETS questions test scores on the rare occasions when concerns arise about their validity. Fairness — to test takers and the colleges, universities, and others that rely on ETS test scores in making important decisions about test takers — requires ETS to review test scores that may be invalid. For test scores to be valid, they should be earned under standard conditions, and no test taker should have an unfair advantage.

Almost all test scores are reported by ETS without questioning their validity. Only about one-tenth of one percent of test scores is ever questioned by ETS. When we must question a test score, we encourage the test taker to submit information that addresses our concerns. We also make available four options for resolving the matter. The test taker may:

1. Take a free retest to confirm the questioned score; *or*
2. Authorize ETS to cancel the questioned score; *or*
3. Let the university (or other institution) decide; *or*
4. Submit the matter to arbitration at ETS's expense

All of these options (the test taker also has the right to challenge ETS's decision in court) are detailed in this booklet, which we have prepared in the spirit of ensuring the highest quality assessments and results for everyone concerned. The procedures described in this booklet do not apply to certain group irregularities, cases of reported misconduct, or licensing and certification tests.

First, Why We Question Test Scores

For about 50 years, ETS has developed, administered, and scored standardized educational tests with two goals in mind: quality and fairness. ETS tests are widely viewed as accurate assessments of the abilities they are designed to measure. As a result, the millions of people who take ETS tests each year and the thousands of institutions that receive test score reports have come to rely, and justifiably so, on the validity of test scores reported by ETS. That is why ETS must question test scores when we believe there is substantial evidence that they are not valid.

Mutual Agreement

ETS is required by contract to administer tests under secure, uniform conditions that afford all test takers the same opportunity to demonstrate their abilities. When people register for an ETS test, they agree to accept our procedures. They also acknowledge that ETS has the right to review scores and to question and cancel scores when we believe there is substantial evidence that they are invalid.

The Score Review Process

Over the years, ETS has developed procedures to review the validity of test scores. Unless ETS finds substantial evidence that the scores are invalid, test scores are reported. When ETS finds there is substantial evidence that a score is invalid, we notify the test taker and offer a menu of options for resolving the matter. If it cannot be resolved, ETS must cancel the questioned score.

Assuring Score Validity

ETS does not base its decision to question or cancel a test score on a finding that any test taker cheated on the test. In most cases, it is neither ETS's intention nor responsibility to make judgments about whether a test taker cheated, and we do not express such judgments to third parties. Our responsibility is to assure, based on information available to us, that test scores we report are valid to the extent possible.

Fairness and Privacy Safeguards

ETS recognizes the importance of treating test takers fairly — and we have designed our procedures with fairness in mind. Our communications in cases of questioned scores — including this booklet — are designed to help test takers understand our procedures for reviewing scores so questions about score validity can be resolved as quickly, economically, and equitably as possible.

ETS strives to protect the privacy of test takers whose scores are questioned. This means we avoid discussing with anyone facts that would identify a test taker — unless the test taker has made public filings or statements concerning his or her questioned scores to which ETS believes it has a right or obligation to respond. Otherwise, ETS discusses personal information only with persons designated by the test taker.

Advice from Others

ETS recognizes that test takers may seek advice from persons they believe can be helpful. Minors and high school students may, of course, talk to their parents, teachers, guidance counselors, and others who can be of assistance. In addition, Test Security staff are available to discuss these matters with the test taker or anyone who has been asked by the test taker to help resolve these questions.

How Questions Arise

ETS receives information about questionable scores from various sources. Among other things, we compare each test taker's current scores with previous test scores, frequently using a "large score difference" measure to identify scores that warrant review. However, ETS never invalidates test scores based on large score differences alone.

Other sources of information include:

- Communications from test center supervisors, proctors, and other test takers;

- Inquiries from colleges, universities, and other score users about the validity of particular scores (such inquiries often arise from inconsistencies among different measures of the test taker's ability); and
- Other miscellaneous sources of information.

Although ETS considers information received from these sources, we do not question scores unless we determine for ourselves that there is substantial evidence of invalidity.

No Action Taken During Review

When questions are raised before a test score has been reported, ETS does not report the score to score users until it has been cleared. On the other hand, if a *previously reported* score is in question, ETS does not notify score users unless and until it has decided to cancel the score after completion of the review and resolution process.

Two-Stage Review Process

ETS will not invalidate a test score without substantial evidence that it is invalid. To ensure fairness, the review process has two separate *independent* stages with different sets of personnel responsible for each.

The Initial Review

The ETS Test Security Office is responsible for the initial review of scores. Test Security staff consider whether, based on information available to ETS, there appears to be substantial evidence of invalidity. In rare instances, Test Security staff call on test center staff to obtain more information. If the Test Security Office determines that there is not substantial evidence of invalidity, it terminates the review and sends any scores not already reported to the designated score users. Score users who have questioned the scores are advised that the scores have been cleared.

If Test Security staff find substantial evidence that a score may be invalid, they notify the test taker and give him or her *one* opportunity to submit additional information that addresses their concerns. Upon receipt of such information, or expiration of the period for submitting it, Test Security staff refer the case to the ETS Board of Review for evaluation and decision. Test takers are also offered two options for resolving the matter at this stage; those options — retesting and score cancellation — are described beginning on page 14 of this booklet.

Submitting Additional Information — What Might Make a Difference?

Before questioned scores are submitted to the Board of Review, the Test Security Office provides test takers *one* opportunity to submit information addressing ETS's concerns. Test takers are free to provide any information about their test experience that they feel is relevant — and in some cases, this information resolves ETS's concerns. What information might make a difference?

- Other standardized test results or academic records may tend to show that the questioned score is consistent with other measures of the test taker's abilities.
- Authenticated documents written prior to the questioned test administration may serve to address questions about handwriting differences.
- In the case of a physical impairment or other disability (which may account for substantial score differences or apparent handwriting discrepancies), the test taker may submit a doctor's certificate or other relevant information.

ETS's Board of Review considers all such information.

However, we give less weight to information that does not specifically address ETS's questions about score validity. For example, character references or testimonial letters do not explain handwriting differences or unusual agreement between the answers of two test takers.

The Second Stage of Review

The Board of Review is an impartial group of ETS professional staff. Board of Review members do not review scores from testing programs for which they have managerial or administrative responsibility. The Board meets in panels of three to review cases. If even one panel member concludes there is not substantial evidence of invalidity, the review is terminated and the score cleared.

If the Board of Review finds that there is substantial evidence of invalidity, the Test Security Office notifies the test taker. ETS offers four options at this stage, as outlined below. As discussed on page 11 of this booklet, the first two options are also available before the case is submitted to the Board of Review. (Test takers also have the right to challenge the Board of Review decision in court. However, the filing of a lawsuit will not necessarily delay ETS's cancellation of questioned scores.)

Options Leading to Resolution

Option 1 — Retake the Test at No Charge

One option test takers have for verifying the validity of a score is to take the test again free of charge at a specially arranged administration to confirm that the original score accurately reflects their ability. The new score from a retest has only to be reasonably close to the original score to confirm its validity, in accordance with the statistical guidelines of each testing program. Test Security staff are available to advise test takers of these confirmation standards. If the score on a retest is higher, that score is the one that counts. This option is available only to test takers in the United States and Canada.

Requirements for Retaking the Test

The retest, which ETS arranges as quickly as possible, is administered under secure conditions. Test takers must present positive identification before retesting begins. ETS requires a personal photograph of the test taker, the names of three responsible people who can identify the test taker from the photograph, and a thumb print.

If Validity Is Not Confirmed

If retest scores do not confirm the validity of questioned scores, ETS will cancel the questioned scores and notify any score users to which the scores have already been reported that the scores have been canceled. Test takers may elect, after being informed of their retest scores, to have their retest scores reported to score users instead of the questioned scores. ETS does not refund test fees when test takers choose to report the retest scores.

Option 2 — Test Taker May Cancel the Score

A test taker may authorize ETS to cancel the questioned score. ETS then removes the score from the test taker's record and, if the score has been previously reported, notifies score users that it has been canceled because ETS no longer considers it valid. In such instances, ETS does not disclose the specific reasons for canceling the score and will refund any test fees paid by the test taker.

Option 3 — Let Score User Decide

Test takers may authorize ETS to send the questioned score, along with a summary of their ETS Test Security file, to a college, university, or other designated score user so that the institution can make its own decision about accepting the score for its own use. ETS will not send a summary, however, unless the user agrees to accept it and to protect its confidentiality.

Factors to Consider

Before electing Option 3, test takers should consider three factors:

1. ETS will cancel the score before releasing the file to any score users.
2. Not all score users are willing to review files and make independent judgments about questioned scores.
3. User judgments are not binding on ETS.

Option 4 — Arbitration

A test taker may ask to have a third-party arbitrator, appointed by the American Arbitration Association, determine whether ETS has substantial evidence to support cancellation of the questioned test score. This option is available only to test takers in the United States.

Requirements

Test takers electing this option must sign an ETS Arbitration Agreement that spells out the procedures that will apply in the arbitration. Arbitration is intended only as an independent review of ETS's decision that there is substantial evidence to support cancellation. As a result, the arbitrator will review only the information available to the ETS Board of Review when it decided to cancel the scores. Therefore, test takers may not present evidence in the arbitration that was not submitted to ETS within the time provided.

Cost

ETS usually pays the cost of arbitration. However, the arbitrator may charge the test taker up to \$200 if the arbitrator determines the test taker's position to be frivolous.

Questions About Options

ETS invites test takers to call the Test Security Office about these options. A staff member will explain them and describe their implications in greater detail.

Cancellation Procedures

When ETS cancels a score, it is removed from ETS's files, and any fee the test taker paid is refunded. If a score has not already been reported to any score users, ETS takes no further action. If scores *have* been reported, ETS notifies the designated test score users that the score has been canceled. The specific reason for canceling the score is not disclosed.

Note:

The options described in this booklet do not apply in cases of group irregularities that raise concerns about score validity or in cases of reported test taker violations of test administration rules.

Some Types of Preliminary Information that May Lead ETS to Question Scores

Information that:

- a test taker may have copied or received help from another test taker
- test scores are inconsistent with other measures of the test taker's abilities
- questions or answers may have been available before the test administration
- irregularities occurred during the test

Comparisons of:

- one test taker's answers with those of others
- a test taker's scores with others he or she earned on this test
- the handwriting on the answer sheet with handwriting on other documents
- the information on an answer sheet with other records
- changed answers on the answer sheet with the answers of another test taker
- a test taker's Photo File Record with other records

Overview Guidelines

Educational Testing Service is committed to ensuring that its tests and publications acknowledge the multicultural and multiethnic nature of our society and reflect a thoughtful and fair consideration of the very broad character of ETS's clientele. As part of the effort to attain this goal, ETS has stated in its *Standards for Quality and Fairness* that all ETS products and services — including individual test questions, tests as a whole, and publications in print and other media — must not contain language, symbols, words, phrases, or examples that are generally regarded as sexist, racist, or otherwise potentially offensive, inappropriate, or negative toward any group.

Language

ETS is committed to developing tests, publications, and other materials that are free of racist, sexist, or otherwise potentially offensive language and images.

Stereotyping and Language Use

ETS material will not use language that stereotypes a population subgroup unless the purpose of the material is to examine stereotypes.

The sensitivity reviewer ensures that tests and other materials do not contain language or symbols that may reinforce stereotypes, such as stating or implying that a population group is biologically or culturally superior or inferior to any other group.

In certain areas, such as content tests or research materials, stereotypes and stereotyping may have a legitimate role. When such material is used, it should be handled in a conscientious, balanced, sensitive, and objective manner.

Inflammatory or Highly Controversial Material

ETS material will not contain inflammatory or highly controversial topics (e.g., abortion, euthanasia), except where such material is both relevant and essential. Material that may have negative emotional impact for some population subgroups may have an appropriate role in certain areas, as in research products or content or mixed tests designed primarily to measure knowledge specific to that material. If the material must be used, it should be handled in a conscientious, balanced, sensitive, and objective manner.

Inappropriate Tone

ETS material will not contain language that is inappropriate in tone. A patronizing, insulting, elitist, or inflammatory tone is unacceptable unless the express purpose of the material is to examine such a tone.

Recognition of Population Diversity

ETS is committed to developing tests and other publications that reflect a thoughtful and fair consideration of the diversity of ETS's clientele.

Racial/Ethnic Balance

Wherever possible, no racial/ethnic group should be represented to the exclusion of others in ETS material. Wherever feasible and appropriate, ETS material will include content that highlights or refers to the achievements and contributions of different cultures and racial/ethnic groups.

Gender Balance

ETS products should reflect an appropriate balance of males and females.

In judging balance in tests and publications, the sensitivity reviewer should consider both the balance of gender and minority representation and the overall impression made by the references to women and minority groups.

For test material, the applicability of balance guidelines depends on the classification of a test and the specifications for the test.

All text and artwork as well as video and audio stimuli must be reviewed for conformity to balance guidelines.

The concept of numerical balance may not be applicable to much of the material produced in the Research divisions. Balance is required for such things as the gender of people used in examples and in tests constructed specifically for research.

Test Fairness

ETS is committed to avoiding the inclusion of content that might unfairly disadvantage test takers from any population subgroup.

Elitism, Ethnocentrism, and Inappropriate Underlying Assumptions

ETS products will not contain inappropriate underlying assumptions, in particular, ethnocentric, elitist, and/or gender-based beliefs and language that are not germane to the domain being tested.

Elitist or ethnocentric concepts, words, or phrases should not be included in ETS products. Expressions or words that are likely to be more familiar to members of one social class, religion, or ethnic group are generally unacceptable unless definitions are provided.

Review Procedures

ETS is committed to ensuring that the sensitivity review guidelines are applied consistently and equitably to all ETS products, whether written, visual, or oral.

Material to Be Reviewed

All ETS products will receive a sensitivity review.

Each division at ETS must ensure that its test materials and publications — including instructional packages, test kits, study manuals, and video stimuli for tests — receive timely sensitivity reviews.

Tests

All tests, as well as associated audio or visual stimuli, must receive a sensitivity review. Additionally, each test in use must undergo a sensitivity review every five years. Item banks owned by clients but developed by ETS must also be reviewed.

There are two major components to the sensitivity review process for tests: an optional preliminary review and a mandatory final review.

The optional preliminary review may be conducted at any time during the process of assembling a test and/or instructional material to check for potential problems. Preliminary reviews are particularly recommended for audio and visual stimuli because of the time and expense involved in producing them.

The mandatory final review takes place during the test editing process. Even if a test has received a preliminary review, the mandatory final review must be performed at this time.

Publications

All ETS publications in every medium — print, video, audio, and software — intended for audiences of 50 or more must receive a sensitivity review. ETS materials intended for an internal audience — such as memoranda, planning schedules, budgets, and newsletters — must conform with the sensitivity review guidelines, as should all ETS correspondence. An author who is uncertain whether material is in conformity with the guidelines should check with a sensitivity coordinator.

The sensitivity review of a publication is carried out when the manuscript has reached final draft stage. However, editors are urged to review copy informally for sensitivity concerns as early in the editorial process as possible.

Research

All reports, including program research reports, research reports, research memoranda, statistical analysis reports, and final reports to external agencies, require a sensitivity review. Associated questionnaires and interview protocols must also be reviewed. Normally these reviews are performed during the content review process.

ETS media products developed in the Research divisions (e.g., software, videos) require sensitivity review as early as possible in the development process.

Reviews of Non-ETS Products

Non-ETS products of concern to the sensitivity review process usually fall into one of two groups:

- Those produced in whole or in part by ETS staff. Intellectual products created by ETS staff that are not published by ETS, such as texts for speeches and journal articles, should reflect the spirit of the sensitivity review guidelines. Reviews are strongly encouraged for products that:
 - a. Refer to population subgroups.
 - b. Are produced in relation to work performed for ETS or may be perceived as representing the views of ETS.
- Those produced by clients with ETS involvement. These materials do not require review. Clients should be made aware of sensitivity review guidelines and encouraged to conform with them when there is a significant ETS role in the preparation, production, or distribution of a non-ETS product.

The Sensitivity Reviewer

The sensitivity review is carried out by a qualified sensitivity reviewer. All ETSers are encouraged to attend sensitivity review training and to review materials informally at all stages.

Sensitivity reviewers of record — those who are qualified to sign off on sensitivity review forms — are members of the ETS staff or consultants who

have completed the appropriate training and the required refresher courses. They should also have participated in any further sensitivity review training necessary in their areas.

Prospective reviewers should receive either test development, publications, or research training, depending on the kinds of sensitivity reviews they will be performing.

Disagreement Resolution Procedures

ETS has established procedures by which disagreements in the implementation of the guidelines can be resolved fairly and in a timely manner.

If sensitivity review comments and recommendations are made and agreement cannot be reached on how to resolve the issues raised, the matter is referred to the area sensitivity coordinator for mediation. Issues that remain in dispute are submitted to line management.

Monitoring the Process

ETS will monitor the sensitivity review process and initiate actions, as necessary, to maintain the highest standards.

Responsibility for monitoring the sensitivity review process in each area lies with the respective sensitivity review area coordinators, their line management, and the sensitivity review steering committee.

ETS STANDARDS FOR QUALITY AND FAIRNESS

ADOPTED BY THE BOARD OF TRUSTEES

 EDUCATIONAL TESTING SERVICE • PRINCETON, NJ

PREFACE

iii

Educational Testing Service (ETS) is strongly committed to the principles of openness in testing, public accountability, quality, and fairness. In October 1981, the ETS Board of Trustees adopted and publicly announced as corporate policy the *ETS Standards for Quality and Fairness*. At the same time, the Trustees directed ETS management to maintain a program for monitoring adherence to the Standards and authorized the appointment of a Visiting Committee of persons outside ETS who were to annually review and report to the Trustees on ETS's adherence to the Standards. These actions by the Trustees are tangible evidence of ETS's commitment as a private, nonprofit educational organization to public accountability and to publicly declared standards by which the organization is prepared to be judged. ETS believes that the Standards contribute significantly to the quality and utility of its programs for those institutions and individuals ETS serves.

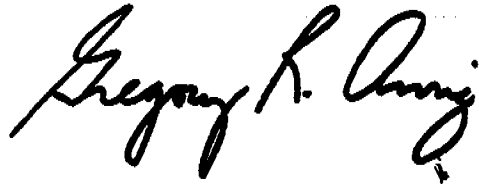
Compliance with these Standards is taken seriously at ETS. The Standards are applied to all ETS-administered programs. Adherence to the Standards is regularly assessed through a carefully structured Audit Program and subsequent management review. The audit is a rigorous process that reviews each program's policies and practices carefully. Management then evaluates every recommendation to assure that appropriate action has been taken.

The ETS Standards and the Audit Program are important matters to the ETS Trustees. To ensure that the Standards are interpreted and applied according to the spirit and purpose intended, the Trustees established the Visiting Committee, which is composed of distinguished educational leaders, experts in testing, and representatives of organizations that have been critical of ETS in the past. The Committee meets annually with ETS staff, senior management, and outside auditors, and it issues a report directly to the Committee on Public Responsibility of the ETS Board of Trustees in June of each year. The Visiting Committee's report is published by ETS and released in its entirety to the media and to any interested members of the public.

The ETS Standards and efforts to apply them reflect ETS's determination to hold itself accountable to high standards of performance and to set high standards for the products and services ETS provides. These efforts have been viewed positively by ETS staff as well as the clients we serve. We take great pleasure in noting the first Visiting Committee's conclusion:

"We find ETS's effort to maintain and improve the quality and fairness of testing well conducted. We know of no other testing organization with anything comparable. The ETS system of auditing its work is an admirable component of ETS's commitment to public accountability; we applaud ETS's intent to be publicly open about activities in which the public clearly has a legitimate interest, even though ETS is a private organization."

This publication represents a continuation of this commitment. In 1985, the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education adopted a comprehensive revision of the *Standards for Educational and Psychological Testing*. The *ETS Standards for Quality and Fairness* were also revised in order to stay in the forefront of measurement and the latest thinking of the profession. These revised *Standards for Quality and Fairness*, which were adopted by the ETS Board of Trustees on June 11, 1987, and the Audit Program through which they are applied are an important part of ETS's continuing commitment to quality, fairness, and public accountability.



Gregory R. Anrig
President
Educational Testing Service

CONTENTS

v

Introduction.....	vi
Accountability.....	1
Confidentiality of Data	3
Quality Control for Accuracy and Timeliness.....	5
Tests and Measurement—Technical Quality of Tests	7
Validity	8
Test Development.....	10
Test Administration	13
Reliability	15
Scale Definition	17
Equating.....	18
Score Interpretation.....	19
Test Use.....	21
Research and Development	23
Public Information	26
Assuring Quality and Fairness	28
The ETS Audit Program	28
The ETS Visiting Committee	30
Glossary.....	31

INTRODUCTION

The *ETS Standards for Quality and Fairness* are designed to ensure that ETS products and services demonstrably meet explicit criteria in seven areas of basic importance: Accountability, Confidentiality of Data, Quality Control for Accuracy and Timeliness, Research and Development, Tests and Measurement, Test Use, and Public Information. The first three sections of the Standards deal with issues that relate to all ETS activities: the responsibilities of ETS to those affected by its activities; the rights to and limitations on access to data collected by ETS; and the control of quality and performance. The remaining sections concern issues relating to ETS's main endeavors: Research and Development, Tests and Measurement, Test Use, and Public Information.

The ETS Standards reflect and adopt the *Standards for Educational and Psychological Testing* jointly issued by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The ETS Standards, however, are tailored to ETS's particular circumstances and needs. Thus, the Standards may not be useful to organizations whose practices, programs, or services differ from those of ETS.

The Standards comprise both principles that underlie ETS efforts in each area and policies that govern decision-making and guide the development of more specific goals. The Standards are implemented by ETS management through procedural guidelines that provide more detailed criteria for ETS's diverse programs and services. The Standards are reviewed and revised from time to time to keep them abreast of developments in professional practice and research.

Like the *Standards for Educational and Psychological Testing* issued by AERA, APA, and NCME, proper interpretation and implementation of ETS's Standards depends on the seasoned judgments of professional staff. These judgments must be carefully based on research, professional experience, and sound reasoning. The ETS Standards are intended to guide and assist ETS professionals in the flexible and sensitive exercise of professional judgment, not to obviate the need for it. Thus, if adherence to any procedural guideline is infeasible or inappropriate in particular circumstances, or if good professional practice in a particular instance conflicts with the letter of a guideline, then sound practice, consistent with the spirit of the underlying principles and policies, should prevail.

ETS does not have sole responsibility or authority to determine how or whether these Standards will be implemented in activities for which practice or policy is substantially established by a group, individual, or institution other than ETS. These Standards are not intended to establish obligations on the part of ETS to act or intervene in situations where the pertinent responsibility rests primarily outside ETS. However, ETS does encourage and assist groups and institutions in implementing the Standards related to any of their activities that involve ETS products or services.

ETS has committed itself to these Standards and to a continuing program of research and development. As a result, ETS expects to expand the realm of knowledge relevant to its activities and to nurture at ETS and elsewhere the development of thoughtful and sensitive professionals with the skills and sensitivity necessary to apply the principles and the policies embodied in these Standards.

CONFIDENTIALITY OF DATA

Principle

ETS recognizes the right of individuals and institutions to privacy with regard to information supplied by and about them that may be stored in files held by ETS and the concomitant responsibility to safeguard information in its files from unauthorized disclosure.

Policies

- A. ETS will ask individuals to provide information about themselves only if it is potentially useful to those individuals, it is necessary to facilitate processing of data, or it serves the public interest by improving understanding of human performance. Individuals should be informed of the purpose for which the information is requested.
- B. ETS will protect the confidentiality of individual and institutional data that may be stored in files held by ETS. This right to privacy extends both to processed information, such as scores based on test-item responses, and to the raw data on which the processed information is based.
- C. ETS will not collect or maintain in files held by ETS any critical information that in its judgment cannot be protected adequately from improper disclosure.
- D. ETS will encourage the organizations with which it works to adopt policies and procedures that adequately protect the confidentiality of the data transferred by ETS to those organizations.

Procedural Guidelines

- 1. Inform individuals or institutions to the extent appropriate, before information is collected, of the information's intended use, the conditions surrounding its confidentiality and release, and the length of time the information will be retained.
- 2. Use or release from ETS identifiable information about an individual or institution only for purposes for which permission has been granted with proper consent or through prior agreement, or in a manner that assures the confidentiality of the individual or institution.
- 3. Permit individuals, on presentation of appropriate identification (e.g., signature and data file number), to authorize the disclosure of information about themselves from ETS-held data files to any appropriate recipient, provided that disclosure does not violate other ETS policies or the privacy of other individuals. If authorization is from a third party by prior agreement with the individual, the individual should be notified when disclosure has taken place.

4. Permit individuals or their legal representatives to obtain information about themselves from data files held at ETS. Such release of information must be consistent with other ETS policies and allowed only upon the individual's submission of appropriate identifying information and, if necessary, payment of a reasonable fee.
5. Assure that access to electronic, paper, or other forms of confidential data is reasonably safeguarded.
6. Develop clear retention guidelines and procedures for eliminating identifiable information from data files.
7. Provide identifiable data only in a manner consistent with these guidelines unless served with a subpoena or other legal process to provide identifiable information. In that event, inform legal counsel in order to make appropriate efforts to narrow the subpoena or to obtain a court order or other arrangements to minimize the dissemination of that information.
8. Inform organizations with which ETS works of the confidentiality of data transferred by ETS to those organizations or collected by them on behalf of ETS so that the confidentiality of such data can be protected.

QUALITY CONTROL FOR ACCURACY AND TIMELINESS

Principle

The accuracy of ETS's products and the timeliness with which they are made available are important parts of the responsibility ETS has undertaken with respect to its sponsors and the diverse public it serves:

Policies

- A. ETS will use adequate quality controls to assure that its products and services are accurate and timely.
- B. ETS will make realistic delivery commitments and reasonable efforts to meet those commitments.
- C. ETS will sacrifice the timeliness of the delivery of information if the desired accuracy of that information is substantially in question.
- D. ETS will seek to inform those negatively affected if, subsequent to its release, information has been found to be inaccurate.
- E. ETS will seek to inform those negatively affected if it is likely that there will be substantial departure from schedule commitments.

Procedural Guidelines

1. Verify that products and services conform to specifications or standards before external release by doing as many of the following as appropriate:
 - independently recomputing or visually inspecting an appropriate sample of each product;
 - assessing the reasonableness of results through reviews by technically competent staff;
 - reviewing and proofing printed material; or
 - assuring adherence to ETS Standards and other professional standards through effective peer review.
2. Verify and document the accuracy of internal products when:
 - the information (e.g., answer keys, conversion parameters, algorithms) is critical to the external product; or
 - early detection and correction of errors would facilitate meeting delivery schedules of the external products.

3. Monitor the accuracy, timeliness, and responsiveness of replies to inquiries.
4. Report to a specified ETS staff member all instances in which a product or service failed to conform to accuracy or timeliness requirements. Resolve discrepant conditions before release of the product unless the cognizant ETS officer has approved release to benefit the majority of product users. 3
5. Correct any critical information found to be in error after its release, and promptly distribute corrected information to those negatively affected by the error.
6. Make provision for individuals to verify scores or other information about them within a reasonable time. Such requests must be accompanied by appropriate identifying information and, if necessary, a reasonable fee.
7. Establish schedules or other process-control methods to assure the timely production of each product or service. If it is likely that a product will be late, take steps (e.g., proper notice to test users) to minimize the effect.

TESTS AND MEASUREMENT— TECHNICAL QUALITY OF TESTS

7

This section, which deals with ETS testing activities, presents related procedural guidelines in seven subsections devoted to validity, test development, test administration, reliability, scale definition, equating, and score interpretation.

Principle

High standards of quality and fairness in constructing, administering, reporting, interpreting, and evaluating ETS tests are central to valid inferences, decisions, and descriptions derived from test scores, and to ETS's capability to function effectively as a test developer and educational service and research organization.

Policies

- A. ETS will develop tests in which the knowledge, skills, abilities, or personal characteristics measured, procedures followed, and criteria used will be appropriate to the use for which the test is designed.
- B. Recognizing that test validation is the joint responsibility of test users and test developers, ETS will encourage and assist test users in their validation efforts and will design tests to meet professionally acceptable standards of validity for their intended use(s).
- C. ETS will strive to develop tests that will be unbiased with regard to major population subgroups being tested.
- D. ETS will establish standard processes for test administration that minimize variations in test performance due to circumstances or conditions not relevant to the attributes being measured.
- E. ETS will construct tests that are sufficiently reliable for their intended use(s).
- F. ETS will develop scales for reporting scores in a rational fashion consistent with the requirements and intended use(s) of the test.
- G. ETS will assure comparability of scores that are derived from different editions of the same test and used to compare individuals or groups.
- H. ETS will make available to score recipients data for interpreting scores on ETS tests that encourage appropriate use of those scores.
- I. ETS will adhere to appropriate professional standards, such as those published in *Standards for Educational and Psychological Testing* and *Principles for the Validation and Use of Personnel Selection Procedures*.

Procedural Guidelines: Validity

1. Describe the construct to be measured, and provide validity evidence related to the intended use(s) of the test scores. Describe how the mix of evidence provided is appropriate for the inferences that are to be drawn and the actions that will result from test scores.
2. Describe validation procedures and the results, including, as appropriate:
 - the logical and empirical analyses of processes underlying performance on the test;
 - the relationships between test scores and other pertinent variables, including likely sources of construct-irrelevant variance;
 - how the test and test items were derived from and are related to a domain of knowledge or skills appropriate to the intended inferences that are to be made;
 - the logical and empirical evidence supporting the discriminant validity of subscores;
 - the number and qualifications of any experts who made judgments, and procedures used to arrive at judgments pertinent to the validation effort;
 - the materials surveyed, and the rationale and procedures for defining test content;
 - the systematic analysis of job functions and the link from job tasks to knowledge, skills, and abilities as well as to test content and items;
 - the rationale and procedures for determining criterion relevance, the selection procedures for and the composition of the validation sample, the relationship between predictors and criteria, and factors that affect the relationship, including technical quality of the criteria (e.g., their reliability, the elapsed time between test administration and criterion data collection, and rules for combining criteria if several are combined); and
 - information relative to the interpretation of quantitative evidence, such as associated standard errors of estimate, adequacy of the sample, possible restriction of range of scores on the variables, unadjusted coefficients (when statistical adjustments are made), the need for cross-validation, and other contextual factors.
3. Base validity evidence in a particular situation (e.g., institution, department, or job study) on data from other situations only when it can be established that the particular situation is from the same population of situations. Include information in the documentation about the similarity of the groups tested, the curricula, the job tasks, or other appropriate criterion variables.
4. Undertake new validity studies whenever there is substantial change in the test, the mode of administration, the characteristics of the intended test-taking population, or the performance domain sampled.

5. Whenever there are sufficient population subgroup members to permit meaningful analyses, investigate validity for major subgroups if the need for such investigation is indicated by consideration of the intended use(s) of the test scores, the characteristics of the intended test-taking population, or research.
6. Establish test names that imply no more than the validity evidence justifies.
7. Provide information to users to help them plan, conduct, and interpret validity studies with respect to the intended uses of the test scores such as pass/fail decisions, selection, or placement.

Procedural Guidelines: Test Development

1. Obtain substantive contributions to the test development process from qualified persons who are not on the ETS staff and who represent relevant perspectives, professional specialties, population subgroups, and institutions. Document their relevant qualifications and characteristics.
2. Ascertain basic information for each test to be developed, including:
 - the test's intended use(s), including inferences to be made from test performance;
 - the intended population that will take the test, including anticipated major subgroups; and
 - a description of the underlying construct and the procedures followed for defining the domain to be assessed, a description of the domain, and a description of its relevance to intended test use(s).
3. Document rationales and procedures for the test being developed, including:
 - the rationale for the item type(s) and test format to be used and whether any background or prior experience factors (e.g., age, linguistic or cultural background of intended test takers) affected item-type or test-format selection;
 - the procedures followed for generating test content to represent the domain or to link test and job content;
 - the rationale for the scoring method(s), especially when judgmental processes are used;
 - the item response model, calibration procedures, and the nature of the sample used to estimate parameters when item response theory procedures are used to assemble the test;
 - the rationale and procedures for making branching decisions, for terminating the test, and for scoring the test when adaptive or branching tests are used; and
 - the rationale or evidence supporting comparability when multiple methods for presenting items or recording responses (e.g., foreign language translations, computer testing, or tests modified for individuals with handicapping conditions) are intended to be used. Establish interpretative guidelines for multiple methods where comparability is not supported.
4. For each test, prepare, with appropriate advice and review, test development specifications that cover the following:
 - Content and Skills—a clear description of what is to be tested, including, where appropriate, critical content to be included in each form, and the relative weight to be given to each part of the domain that is to be measured;
 - Test and Item Format—item types to be used; special requirements regarding directions and sample items or tests;

- Psychometric Characteristics—the intended level of difficulty of the test; the number of items, requirements regarding the target distribution of item difficulties, requirements regarding the homogeneity of items within each test or subtest and the correlation between subtests or tests, requirements for equating, including the content and statistical specifications for equating items, and the testing time allotted or suggested;
 - Sensitivity—requirements for the inclusion of material reflecting the cultural background and contributions of major population subgroups; and
 - Scoring—the procedures for scoring, especially when judgmental processes are used.
5. Assure that time requirements are consistent with the test's purpose so that time is not a decisive factor in performance for the large majority of test takers, except for tests designed to measure rate of performance.
 6. Have subject matter and test development specialists who are familiar with the specifications and purpose of the test and with its intended population review the test items for accuracy, content appropriateness, suitability of language, difficulty, and the adequacy with which the domain is sampled.
 7. Review individual items, the test as a whole, directions, and descriptive materials to assure that:
 - appropriate technical standards such as those contained in ETS item writers' manuals are met;
 - language, symbols, words, phrases, and content that are generally regarded as sexist, racist, negative toward population subgroups, or otherwise potentially offensive are eliminated except when judged to be necessary for adequate representation of the domain;
 - editorial standards for clarity, accuracy, and consistency are met;
 - clear and complete directions appropriate to the nature of the test and the characteristics of the test takers are provided;
 - typography, format (e.g., test book, computer display, tape), test-book layout, and response method do not hinder the task of test takers; and
 - sufficient sample questions and answers are contained in program publications to be representative of test content, item types, and difficulty.
 8. Evaluate the performance of individual items by pretesting, pilot testing (especially for examinees with handicaps), reviewing the results of administering similar items to a similar population, or conducting preliminary item analysis before scores are reported.
 9. Whenever there are sufficient population subgroup members to permit meaningful analysis, use data on item performance relative to subgroups to enhance the judgments of test developers if the need for such studies is indicated by consideration of the recommended use(s) of the test, the characteristics of the intended test-taking population, or prior research.

10. Evaluate the performance of each test edition by:
 - carrying out timely and appropriate item and test analyses, including analyses for reliability, intercorrelation of sections or parts, and speededness; and
 - comparing the test's characteristics to its psychometric specifications.
11. Review periodically the adequacy of fit of item response models and the sample used for estimating item parameters, when item response theory procedures are used to develop, score, or equate the test.
12. Review test content and test specifications periodically to assure their continuing relevance and appropriateness to the domain being tested.
13. Review periodically all active test editions developed in prior years and their descriptions in publications to assure the continued appropriateness of both content and language for the present test-taking population and the subject-matter domain.
14. Analyze major changes in test specifications to assure that they are followed by appropriate consideration of the implications for score comparability and to determine whether test name changes or other cautions to test users about comparisons with earlier tests are necessary.

1. Before the test is administered provide prospective examinees (and, in some programs, parents or guardians as well) with information about the following, as appropriate:
 - the test's intended purpose and what it is designed to measure, typical test items, clear directions for the test and the response method to be used, a description of how scores are derived including formation of composite scores, strategies for taking the test (e.g., guessing and pacing), whether the test contains items not intended to be scored, and the background and experience relevant to test performance;
 - the program procedures and requirements, including test dates, test fees, test center locations, special testing arrangements for handicapped persons or others, test registration, score reporting, score cancellation by examinees or by ETS, or the sponsor, and registering complaints; and
 - test administration procedures and requirements, including those related to identification and admission to the test center, materials permitted in or excluded from the testing room, and the consequences of misconduct.
2. Establish test centers that are convenient, nondiscriminating, comfortable, and accessible to all individuals, including those with disabilities. Locate test centers in both minority and majority communities to foster accessibility.
3. Advise test center staff of the need to minimize distractions and to make examinees comfortable in the testing situation. Instruct staff to be sensitive to the psychological as well as physical needs of examinees. Direct supervisors to consult with or include on the test center staff, when appropriate, subgroup members and persons knowledgeable about handicapping conditions.
4. Provide test center staff with a description of the program, the expected candidate population, the duties of staff, and the procedures for:
 - receiving, storing, and distributing test materials to examinees, and returning them to ETS;
 - admitting examinees to the test center, including ID requirements;
 - administering the test to examinees, including handicapped individuals;
 - using appropriate seating plans and assignments, and monitoring the testing room to reduce opportunities to obtain scores by questionable means;
 - handling of suspected cheating, misconduct, or emergencies; and
 - reporting irregularities (e.g., disturbances, mistimings, defective test questions or materials, power failures, or misconduct) so that, after review, appropriate action can be taken.

5. Provide test center staff with directions (to be read aloud before the test begins) that cover the recording of answers on answer sheets or via other devices, timing of test sections and breaks, guessing strategies, and the consequences of using unauthorized aids or engaging in other forms of misconduct.
6. Utilize effective and equitable procedures for preventing, identifying, and resolving scores obtained by questionable means.
7. Encourage examinees to report any irregularities so that, after review, appropriate action can be taken.
8. Undertake quality control activities (e.g., test center observations, solicitation of suggestions from test administrators and examinees, training of test administrators) to assure effective and, when necessary, secure test administrations.
9. Make tests available at no additional cost to individuals with handicapping conditions through special testing arrangements or special test editions.
10. Provide users of locally administered tests with instructions about standardized conditions for administering and scoring the tests.

Procedural Guidelines: Reliability

15

1. Assure that test scores, including subscores and combinations of scores, are sufficiently reliable for their intended use(s). Provide information on reliabilities, standard errors of measurement, or other equivalent information (e.g., information on classification consistency) so that test users can also judge whether reported test scores are sufficiently reliable for their intended use(s).
2. Provide test users with information about sources of variation (e.g., test form, content, population of readers, time interval between testing, practice, and other sources of error) considered significant for score interpretation.
3. Estimate the reliability or consistency of reported test scores by method(s) that are appropriate to the nature and intended use(s) of the test scores and that take into account sources of variance considered significant for score interpretation.
4. Document the reliability analysis, including:
 - a description of the method(s) used to assess the reliability or consistency of the test scores and the rationale for using them, the major sources of variance accounted for in the reliability analysis, and the formula(s) used and/or appropriate references;
 - a reliability coefficient, an overall standard error of measurement, an index of classification consistency, or other equivalent information about the consistency of the test scores;
 - standard errors of measurement or other measures of score consistency for score regions within which decisions about individuals are made on the basis of test scores;
 - the degree of agreement between independent scorings when judgmental processes are used;
 - the adjusted and unadjusted coefficients if reliability estimates are adjusted for restrictions of range;
 - correlations between short forms of tests and the standard form;
 - speededness data; and
 - correlations among reported subscores within the same test or the scores within a test battery.
5. Describe the conditions under which the reliability estimates were obtained, including:
 - a description of the population involved (e.g., demographic information, education level, employment status);
 - the selection procedures for and the appropriateness of the analysis sample, including the number of observations, means, and standard deviations for the analysis sample(s) and any group(s) for which reliability is estimated;

- when scores are based on judgments, the basis for scoring, including selecting and training scorers, and the procedures for allocating papers to scorers and adjudicating discrepancies; and
 - the time intervals between testings, the rationale for the time intervals, and the order in which the forms were administered if alternate-form or test-retest methods were used.
6. Whenever there are sufficient population subgroup members to permit meaningful analysis, study the reliability or consistency of reported scores for major subgroups if the need for such studies is indicated by consideration of the intended use(s) of the test, the characteristics of the intended test-taking population, or prior research.

Procedural Guidelines: Scale Definition

17

1. Establish scales for reporting scores that are well-constructed throughout their range and in a way that facilitates meaningful score interpretation relative to intended use(s) of the scores. Describe the characteristics of the score scale.
2. Establish scale values to be reported that do not encourage finer distinctions among test takers than can be supported by the precision of the test.
3. Choose the scale values in a manner that avoids confusion with other scales that are widely used by the same population of score recipients.
4. Describe the rationale for and the methods used to determine score scales, including:
 - If scores derived from different tests in a program are to be directly compared, take into account, in establishing the scale(s), the differences among groups taking the different tests.
 - If the scale is to be normative, consider the probable length of time and the extent to which the normative information will be appropriate and useful for the intended population.
 - If a test or test battery yields multiple scores for an individual and comparisons among scores are encouraged, establish scales in a manner that allows meaningful comparisons among scores (e.g., normatively or against an absolute standard), or provide data to allow such comparisons.
 - If the scale is to be defined with reference to performance standards, classifications, or cut scores, document the method and rationale used, and the qualifications of any judges.
 - If a scale is used to report composite scores derived from weighting subscores, clearly state the rationale and the method for weighting the subscores.
5. Avoid reporting raw scores or percentages of questions answered correctly on a test or subtest except under one or more of the following circumstances:
 - only one edition of the test is to be offered;
 - scores on one edition will not be compared with scores on another;
 - raw scores on all editions are comparable; or
 - raw scores are reported in a context that supports the intended interpretation(s).
6. Report item responses for individuals or groups only in a context that supports the intended interpretation(s).
7. Reexamine score scales periodically. Redefine an established scale only under compelling circumstances. Provide announcements to all score recipients indicating the change and cautioning recipients against comparisons with earlier scores. If the numerical values are to be changed, change them substantially to minimize confusion between the old scale and the new one.

Procedural Guidelines: Equating

1. Establish equating procedures with the highest level of precision practicable when scores on different test editions are intended to be comparable.
2. Describe the methods used to achieve comparability, including:
 - the rationale for selecting the methods used;
 - the consistency between the assumptions underlying the method and the circumstances under which the method is applied (e.g., when test editions are equated using common items, make the directions, context, speededness, item placement, and other aspects of the test as nearly the same as possible for all examinees; when anchor scores are based on a test that is not representative of the tests being equated, make sure the groups of examinees used for equating are equivalent; or when item response models are used, make sure that information is presented on the adequacy of fit of the model to the data);
 - the procedure for adequately linking all editions of the test for which scores should be comparable; and
 - the plans for specially designed studies to collect data to achieve comparability if only a limited number of editions are offered to institutional or other users who will administer and score the tests.
3. Report the results of the equating experiment, including:
 - the nature of the population involved;
 - a description of the analysis sample(s), including the number of observations, means, and standard deviations;
 - the time intervals between testings; and
 - other statistics appropriate to the method used (e.g., correlation between the anchor test, if used, and the total test).
4. Periodically assess the results of methods used to achieve comparability of scores and evaluate the stability of the score scale.

Procedural Guidelines: Score Interpretation

19

1. Provide score interpretive information for all score recipients in terms that facilitate appropriate interpretations. Provide information that is appropriate for each category of score recipient (e.g., examinee, teacher, college, agency, or media) and that minimizes the possibility of misinterpretation of individual scores as well as group results.
2. Provide each category of score recipients with appropriate information that:
 - describes the intended use(s) of the test and what it is designed to measure;
 - recommends only those score interpretations for which supporting information is available;
 - describes scale properties that affect score interpretation and use;
 - explains the variability of and limitations on the accuracy of test scores (e.g., standard error of measurement, classification errors), and encourages recipients to take such information into account in making decisions based on scores;
 - supports assessments based on individual items or clusters of items whenever such uses are suggested;
 - gives the minimum score(s) required to pass the test when results are reported as pass/fail and examinees have failed the test, and gives general information about an examinee's performance relative to the minimum score(s); and
 - indicates scores on achievement tests might be improved with short-term study, whereas improved scores on aptitude tests might require longer-term preparation.
3. Provide score recipients with an appropriate frame of reference for evaluating the performance represented by test scores through information based on norms studies, carefully selected and defined program statistics, logical analysis, or special studies. When statistical information is included, the information should be adequately labeled and the nature of the group(s) on which the information was based should be clearly identified.
4. Caution score recipients about:
 - scores for different tests offered by a program that may not be comparable even though the scores are reported on similar scales;
 - interpretations that have not been adequately validated (e.g., ones based on foreign language translations, untimed tests for handicapped persons, experimental tests);
 - interpretations based on insufficient numbers of cases;
 - scores that may no longer be comparable because test content or specifications have changed significantly;

- scores earned in previous years that have become of limited value due to changes in the individual or the meaning of test scores over time; and
 - decisions based on the differences between test scores for an individual (e.g., mathematical computation and mathematical reasoning) that do not take into account the overlap between the constructs and the reliability of the score difference.
5. Provide score recipients with information as appropriate to assist them in using scores in conjunction with other information, setting cut scores, interpreting scores for major subgroups, conducting local norms studies, and developing local interpretive materials.
 6. Develop score interpretive information by appropriate method(s) (e.g., norms studies, derivation of program statistics, cut score studies). Describe the method(s), including relevant information about:
 - the characteristics of the scale and procedures used to maintain it;
 - the method of selecting participants on whom data are based, including information about representation of relevant major subgroups within the defined population;
 - the participation rate of categories of individuals or institutions and their characteristics such as the age, sex, or subgroup composition of the group; weighting systems or other adjustments made to form the norming sample; and whether or not the participants were self-selected;
 - the period in which the data were collected;
 - appropriate group statistics whenever tests are intended to be used to make assessments of such groups (e.g., classrooms) rather than individuals;
 - methods and rationale for aggregating test results or developing composite scores;
 - estimates of sampling error and possible effects of nonparticipation;
 - comparisons with relevant data on variables from other sources when possible; and
 - evidence supporting the cut scores or configural scoring rules when different score interpretations are automatically provided for examinees scoring at different points on the scale.
 7. Revise norms or other score interpretation information at sufficiently frequent intervals to assure its continued appropriateness as a frame of reference for evaluation of performance represented by test scores.
 8. Periodically monitor the participation of major population subgroups and, where sample sizes are sufficient, analyze their performance.
 9. Avoid developing interpretive information for population subgroups unless sufficient data are available on each subgroup to make the information meaningful, the information can be accompanied with a carefully described rationale (e.g., guidance purposes) for using it, and the information can be presented in a way that discourages incorrect interpretation and use.

Principle

Proper and fair use of ETS-developed tests is essential to the social utility and professional acceptance of ETS's work.

Policies

- A. ETS will set forth clearly to all score recipients the principles of proper use of tests and interpretation of test results.
- B. ETS will establish procedures by which fair and appropriate test use can be promoted and misuse can be discouraged or eliminated.

Procedural Guidelines

1. Provide score recipients (e.g., examinees, teachers, colleges, agencies, or the media) with adequate descriptions of intended test use(s), caution them about making interpretations not supported by validity evidence, and warn them against reasonably anticipated misuses.
2. Encourage test users to put test scores in an appropriate perspective (e.g., test scores used for selection should be augmented with other relevant information about the examinee for proper interpretation of the abilities being tested; examinees needing to demonstrate critical knowledge or skills for certification should be given multiple opportunities to retest or to demonstrate those skills by other valid and reliable means).
3. Provide users with opportunities for consultation about test use and with information about validity, reliability, test content, test difficulty, and representative research.
4. Advise users that when using test scores differently for members of different population subgroups (e.g., using separate sex norms or using racial data in regression equations), such uses should be carefully and rationally supported.
5. Advise users that whenever individuals are assigned to groups on the basis of test scores, users should undertake periodic examinations of:
 - pass-fail or cut-score policies;
 - the rationale and methods for making assignments;
 - the performance of individuals within their respective groups, where feasible, including the collection of empirical evidence to support the assignments;
 - the continued appropriateness of assignment criteria; and
 - classification rates across major population subgroups.

6. Investigate complaints or allegations of improper score use. When a misuse is verified, advise the sponsor and the user and seek voluntary correction. If efforts to achieve voluntary correction are not successful, consult with the sponsor to determine whether to continue services to the misuser. Maintain records of complaints and their disposition.
7. Assure the accuracy of any ETS-produced promotional material¹ concerning tests and their intended uses.

Principle

A continuing program of research and development conducted in compliance with professional standards with respect to quality and ethical procedures is necessary to maintain the high quality and social utility of ETS's contributions to education and society. This includes basic inquiry to increase understanding of educational processes and human development; public policy and applied research in response to the needs of the educational community, the workplace, and society at large; and research and development to improve ETS products and services. Publication of the results of significant ETS research is of benefit to ETS and the profession because it permits others to use, build upon, or improve ETS work.

Policies

- A. ETS will devote appropriate research efforts to the following:
- Improving measurement and education through the discovery and conceptual integration of new principles and understanding. This research will be aimed at extending knowledge of measurement principles and practices, of the learner and learning processes, of learning environments and educational treatments, of educational institutions, and of the interacting factors that influence human development.
 - Improving the technical quality and the utility of ETS products and services. Among the important issues addressed by this research will be problems of validity, test development, reliability and generalizability, equating, and the soundness of test score interpretation.
 - Responding to the measurement and educational needs of society, and creating, improving, and evaluating instruments, systems, and programs of service that meet these needs.
 - Investigating special problems faced by population subgroups involved in test taking. In addition, ETS will encourage analysis of these groups whenever information about them is pertinent to the research being undertaken.
- B. ETS will conduct its research under appropriate review procedures that protect the rights of privacy and confidentiality of human subjects or respondents and of cooperating institutions.
- C. ETS will follow professional standards and appropriate review procedures to ensure that ETS research is of high quality.
- D. ETS researchers will adhere to appropriate professional and ethical standards, including those published in *Ethical Principles in the Conduct of Research with Human Participants*, *AERA Guidelines for Eliminating Race and Sex Bias in Educational Research and Evaluation*, and *Ethical Standards of Psychologists*.

- E. ETS will encourage the dissemination of full accounts of ETS research in the usual professional forums and will provide other means by which the results of ETS research can be disseminated.

Procedural Guidelines

1. Provide for a periodic assessment of research and development priorities to assure an adequate balance of resources directed toward:
 - improving knowledge of measurement, occupations, educational processes, and human development;
 - meeting the needs of the educational community and society, including population subgroups;
 - improving ETS products and services and the manner in which they are used; and
 - developing new methodologies (including educational, psychometric, and statistical) and technological procedures.
2. Assure the welfare and the right to confidentiality of human subjects or respondents in each project by following procedures approved by the Committee on Prior Review of Research. Procedures approved by the committee include obtaining appropriate informed consent, separating participants' names from data and other steps relating to confidentiality, and avoiding any negative consequences of participation.
3. Report the results of research to participants and institutions with appropriate care so that the possibility of misinterpretation and misuse is minimized.
4. Follow review procedures for research proposals and reports that will assure that research is of high quality. Reviews may include the following considerations:
 - the rationale for the research;
 - the soundness of the design;
 - the thoroughness and care of data collection and analysis;
 - the reasonableness of the interpretation;
 - the clarity of the exposition;
 - the sensitivity to language or material that is generally regarded as sexist, racist, or otherwise offensive or inappropriate; and
 - the soundness of the project planning and management.
5. Publish or otherwise disseminate the results of research projects unless a justifiable need to restrict dissemination is identified before the research begins.

6. Whenever information on sex, ethnic, racial, or other population subgroups is pertinent to the research, studies should be designed to allow analyses of these groups.
7. Provide non-ETS researchers with reasonable access to ETS-controlled nonproprietary data so long as the privacy of individuals and organizations and ETS's contractual obligations can be protected. Grant access to data facilitating the reanalysis and critique of published ETS research with the same requirements for confidentiality of individuals and institutions. Encourage program sponsors and other organizations to adopt similar policies.

PUBLIC INFORMATION

Principle

ETS is dedicated to promoting public understanding of testing, measurement, and related educational issues by providing programs of public information, research, and advisory and instructional activities.

Policies

- A. ETS will promote understanding of the purposes and procedures of testing and the proper uses of test information among examinees, test users, and the general public; ETS will encourage testing program sponsors to undertake similar efforts.
- B. ETS will adhere to high professional and ethical standards in both the promotion and the use of its products and services and in the dissemination of information to examinees, test users, and the general public. ETS will encourage sponsors and other organizations to do so.
- C. ETS will provide instruction and technical assistance in testing, measurement, evaluation, and related areas.
- D. ETS will disseminate the results of research on testing, measurement, and other related educational issues and will make ETS-controlled nonproprietary data available to other researchers; further, ETS will encourage other organizations to do the same.
- E. ETS will respond promptly and appropriately to requests for advice and technical assistance related to: programs and services offered by ETS, purposes and procedures for testing, uses and misuses of test information, and complaints about its products and services.
- F. ETS will collect reference materials relating to tests, measurement, evaluation, and related research, and will make its collections available to professional groups, organizations, and interested individuals.

Procedural Guidelines

- 1. Develop and disseminate publications and other materials, directly and in collaboration with sponsors, that promote proper test use, discourage misuse, and improve public understanding of testing, measurement, and related educational issues.
- 2. Convene periodically groups of test users, measurement specialists, representatives of professional groups, and other interested parties to examine ETS procedures and recommend improvements in them.

3. Provide accurate and appropriate information when describing ETS products and services.
4. Provide advice and technical assistance on tests and measurement for test sponsors, users, and other interested groups.
5. Offer conferences, seminars, workshops, and other forms of training or instruction in testing, measurement, and other relevant areas of interest, acting independently or in cooperation with other institutions or professional groups.

ASSURING QUALITY AND FAIRNESS

The ETS Audit Program

The ETS Standards are but a single part of an overall ETS effort to ensure that its products and services meet appropriate professional standards. Another important part of the process is the ETS Audit Program, administered by the Office of Corporate Quality Assurance, through which the ETS Standards are vigorously applied to all ETS programs.

Every program at ETS must be reviewed relative to the ETS Standards at least once every three years. The audit process has four phases. In the first of these, the area vice presidents select the programs to be reviewed and the Office of Corporate Quality Assurance appoints teams of independent reviewers to assist in conducting the audit. Most testing programs are evaluated by one three-person team on practices related to the Tests and Measurement procedural guidelines, and by a second team on practices related to the other procedural guidelines. Each team's membership is structured to maximize the knowledge and skills necessary to evaluate the program, the diversity of auditor perspectives, and the areas of employment within ETS. Reviewers must be independent of the program they are reviewing. Some audit teams each year also contain individuals not employed by ETS. The Office of Corporate Quality Assurance provides the necessary training and orientation for all audit teams and directors of programs being reviewed.

In the second phase of the Audit Program, program directors evaluate the ways in which their programs have complied with each of the procedural guidelines. They also assemble for the audit teams the program-related information required to verify their evaluation, because the program staff carries the burden of establishing that the program practices are reasonable in light of the ETS Standards. The final step in phase two is a meeting between the program staff and the audit teams. The meeting provides auditors with an overview of the program and the program practices relative to each area of the Standards. Auditors are given an opportunity to ask questions about the program and the documentation provided and to ask for additional information that might help them during their review.

In the third phase of the audit, each team evaluates the program's practices to determine whether or not they measure up to the ETS Standards. Audit teams are instructed to review all program policies and practices—even those that may be the responsibility of the program's sponsor. Auditors are also instructed to be thorough and to determine only whether or not a program practice is measuring up to the intent of the Standards. The circumstances surrounding a possible variance with the Standards are not to be considered. Thus, matters such as technical feasibility or even sponsor acceptance are not considered when an audit team makes its recommendations to the program. Whenever members of an audit team believe that a program does not comply with the Standards, they must explain why and make an appropriate recommendation for resolving the

situation. Each audit team's findings and recommendations are presented at a second meeting with the program staff. Any problems noted during the audit are discussed, as are actions that the program might take to resolve these situations. At this second meeting, program staff can discuss or challenge recommendations made by the audit teams.

When the audits of a program have been completed, the preliminary reports are sent to the program director for his or her review. If, after studying the audit teams' reports, the program director feels that the audits have failed to interpret a procedural guideline correctly, or misstated a program practice in their evaluation, an appeal of the evaluation may be filed with the director of the Office of Corporate Quality Assurance. The director reviews each appeal and notifies the program director of the appeal's disposition. Following any appeals, the final report is submitted to the responsible vice president for follow-up action.

The fourth phase of the audit involves the preparation of reports and the initiation of program actions to address all recommendations in areas where the program was judged not to be measuring up to the ETS Standards. Programs are expected to take appropriate action on every such recommendation. Program directors meet with statistical, test development, or other staff and others to determine the best actions to be taken in light of the audit teams' recommendations. Frequently, program staff will discuss the audit teams' recommendations with the program's sponsor.

In some instances, however, the required action may be unfeasible, unacceptable to the program's sponsor, or even technically impossible. In this instance, the program director might determine that the program should not comply with the appropriate Standards and ask the responsible vice president for an exemption from complying with the procedural guideline in question.

As a final step in the audit process, the Office of Corporate Quality Assurance summarizes the Audit Program results and the actions taken by programs in response to recommendations. This summary includes any appeals submitted by program directors and any exemptions granted by vice presidents. The OCQA summaries are submitted each year for review by senior management, the ETS Visiting Committee, and the Trustee Committee on Public Responsibility.

The OCQA also prepares a summary of program problems of compliance that have continued since the prior audit. These programs are reviewed by ETS senior management to determine whether continuation of the program is appropriate in light of these continuing situations. The decisions of ETS senior management are also reviewed by the ETS Trustee Committee on Public Responsibility.

The ETS Visiting Committee

The first ETS Visiting Committee was formed in 1982 to aid the ETS Trustees in assuring the appropriate application of the ETS Standards. The Committee meets annually and consists of persons from outside ETS and its Board of Trustees who are knowledgeable in the areas addressed by the *ETS Standards for Quality and Fairness*. A special effort is made to include on the Committee individuals who would take a critical, objective view of ETS activities.

Each year the Visiting Committee reviews the Audit Program to assure that audits are appropriately comprehensive and reasonably designed to assess compliance with the ETS Standards. The Committee also determines whether audit recommendations and resulting actions demonstrate an appropriate commitment to the Standards. In addition to the Audit Program, the ETS Visiting Committee reviews other activities, such as program-related research and public information, to assure that ETS activities demonstrate an appropriate commitment to the Standards. The Visiting Committee has unrestricted access to all participants, including auditors, program staff, and management as well as program audit reports and other information relevant to the questions being raised.

The Visiting Committee's report and recommendations are presented directly to the ETS Trustee Committee on Public Responsibility. The report is also issued to the public. The work of the Visiting Committee adds another dimension to the Audit Program and to ETS's continuing effort to live up to the principles of openness in testing, public accountability, and quality and fairness.

- Absolute Standard*—A cut score or performance standard that is established without reference to the score distribution of the people for whom it will be operational. For example, a passing score set at 80 percent of the questions correct without basing the decision on how many people will score above or below that point is an absolute standard. See *Cut Score, Performance Standard*.
- Accuracy*—The extent to which a product conforms to its specifications.
- Achievement Test*—A test that measures a particular body of knowledge or set of skills, usually after training or instruction has been received. Compare *Aptitude Test*.
- Adaptive Test*—A test administered so that the next item presented to a person depends on that person's response to a previous item or set of items.
- Adjusted Coefficient*—A statistic that has been revised to estimate its value under conditions other than those in the sample on which it has been calculated. For example, a correlation coefficient may be adjusted to account for restriction of range. See *Restriction of Range*.
- Alternate Form*—An edition of a test that is written to meet the same specifications and is comparable in most respects to another edition of the test except that some or all of the questions are different. See *Test Specifications*.
- Alternate Form Reliability*—An estimate of reliability based on the correlation between alternate forms of a test administered to the same group of people. See *Alternate Form, Reliability*. Compare *Test-Retest Reliability*.
- Analysis Sample*—The group of people on whose performance a statistic or set of statistics has been calculated.
- Anchor Test*—A short test or section administered with two or more forms of a test for the purpose of equating those forms. See *Common Items, Equating*.
- Answer Key*—A listing of the correct responses to a set of test questions.
- Aptitude Test*—A test that is usually not closely related to a specific curriculum and that is used primarily to predict future performance. Compare *Achievement Test*. Note that the distinction between aptitude tests and achievement tests may depend more on differences in test use than on differences in test content.
- Attributes*—Qualities or characteristics, such as command of a body of knowledge, ability to perform certain skills, or interest in performing a particular type of task.
- Branching Test*—See *Adaptive Test*.
- Classification Error*—(1) The proportion of inconsistent or incorrect categorizations of examinees that would be made on repeated administrations of the same test or of a test and an alternate form, assuming no changes in the examinees' true performance levels. (2) The assignment of an examinee to the wrong category, such as passing a person who lacks minimal competence and should fail.

Classification Rates—The proportions of examinees placed in various categories, such as pass/fail, on the basis of test scores.

Committee on Prior Review—An ETS institutional review board that reviews proposed and ongoing research to ensure adequate protection of human subjects.

Common Items—A set of test questions that remains the same in two or more forms of a test for purposes of equating. The common items may be dispersed among the items in the forms to be equated or kept together as an anchor test. Compare *Anchor Test*. See *Equating*.

Comparable Scores—Scores that are put on the same scale so that they have the same meaning in terms of relative ranking within a defined group of people but cannot necessarily be used interchangeably. For example, percentile rank scores on a reading test and on a math test are comparable scores if the percentile ranks have been based on the same norm group for both tests.

Composite Score—A score that is the combination of two or more scores by some specified formula.

Configural Scoring Rule—A specified procedure for interpreting the pattern of a person's scores on two or more tests or subtests.

Consent—Permission granted by an individual or that individual's parent or guardian for the collection, use, or release of data held by ETS; generally granted upon receipt of a reasonable explanation of the information's intended use and a reasonable explanation of the manner in which the results will be reported.

Construct—A psychological characteristic (writing ability, numerical ability, logical reasoning) considered to vary across individuals. A construct is not directly observable; instead it is a theoretical concept.

Conversion Parameters—Quantitative rules for expressing scores on one test form in terms of scores on an alternate form. See *Alternate Form*, *Equating*.

Criterion—That which is predicted by a test, such as college grade-point average or job-performance rating.

Criterion Relevance—The extent to which the measure used in assessing a test's predictive validity is related to the test's intended purpose.

Critical Content—Knowledges, skills, or abilities that must be measured in a test because of their importance.

Critical Information—Information that will be used to draw important inferences (a) about the sponsor, ETS-appointed external committees, institutional or agency user, examinee, subject or respondent, or (b) by the sponsor, institutional or agency user, examinee, subject or respondent, and which, if incorrect, could be harmful.

Cross Validation—The application of scoring weights or prediction equations derived from one sample to a different sample to allow estimation of the extent to which chance factors determined the weights or equations or inflated the validity estimated in the analysis sample.

Cut Score—A point on a score scale at or above which examinees are classified in one way and below which they are classified in a different way. For example, if a cut score is set at 60, then people who score 60 and above may be classified as “passing” and people who score 59 and below classified as “failing.”

Domain—A defined universe of knowledge, skills, abilities, attitudes, interests, or other characteristics.

Equating—A statistical process used to convert scores on two or more alternate forms of a test to a common scale so that the scores may be used interchangeably. See *Anchor Test*, *Common Items*, *Conversion Parameter*.

ETS Board of Trustees—The ETS Board of Trustees is the governing body of ETS. There are 17 trustees. Sixteen are elected for four-year terms. New members are elected by incumbent trustees. The President of ETS is an *ex officio* member.

Examinee—An individual who takes a test.

Formula Score—Raw score on a multiple-choice test after a correction for guessing has been applied, usually the number right minus a fraction of the number wrong. See *Raw Score*.

Handicapping Condition—(1) A visual, auditory, other physical or learning disability that causes a test administered under standardized conditions to result in a score that significantly underestimates the person's true ability. (2) A disability that limits a person's access to a testing site. See *Standardized Conditions*.

Institutional User—An organizational recipient of information produced or processed by ETS.

Item—A test question.

Item Analysis—A statistical description of how an item performed within a particular test when administered to a particular sample of people. Data often provided are the difficulty of the question, the number of people choosing each of the options, and the correlation of the item with the total score or some other criterion.

Item Response—(1) A person's answer to a question. (2) The answer to a question coded into categories such as right, wrong, or omit.

Item Response Theory—A set of propositions relating people's performance on test questions to certain characteristics of the people and certain characteristics of the items by means of mathematical models. It is based on the assumption that the probability of a correct response by a person to an item can be calculated from the examinee's estimated ability and certain statistical characteristics of the item.

Item Type—The observable format of a test question. At a very general level “item type” may refer, for example, to multiple-choice or free-response questions. At a finer level of distinction, “item type” may refer, for example, to synonym questions or antonym questions.

Local Norms—A distribution of scores and related statistics within an institution or closely related group of institutions (such as the schools in one district) used to give additional meaning to test scores by serving as a basis for comparison.

Locally Administered Test—A test that is given by an institution at a time and place of the institution's own choosing.

Normative Scale—A way of expressing a score's relative standing in the distribution of scores of some specified group.

Parameter—(1) The value of some variable for a population as distinguished from an estimate of the value based on a sample drawn from the population. (2) In item response theory, one of the characteristics of an item, such as its difficulty. See also *Conversion Parameter*.

Part Score—A score derived from a subset of the items in a test. Synonym of *Subscore*.

Performance Standard—A cut score or a defined level of performance on some task. For example, "Run 100 yards in 12 seconds or less." See *Cut Score*.

Pilot Testing—Small-scale tryout of test questions or a test form, often involving observation of and interviews with examinees.

Population Subgroup—A part of the larger population that is definable according to various criteria as appropriate (e.g., by sex, race or ethnic origin, training or formal preparation, geographic location, income level, handicap, or age).

Precision—The width of the interval within which a value can be estimated to lie with a given probability. The higher the precision, the smaller the interval required to include the value at any given probability.

Principles For The Validation and Use of Personnel Selection Procedures, The Society for Industrial and Organizational Psychology, Inc., College Park, MD, 1987.

Raw Score—(1) The number of items answered correctly on a test. (2) In some usages, the formula score is also called a raw score. See *Formula Score*.

Regression Equation—A formula used to estimate the value of a criterion, given the value of one or more observed variables. For example, estimating college-grade point average, given high school grade-point average and SAT scores.

Reliability—An indicator of the extent to which test scores will be consistent across different conditions of administration and/or administration of alternate forms of the test. See *Alternate Form Reliability*, *Test-Retest Reliability*.

Respondent—An individual who provides data to a research project.

Response Method—The procedure used by an examinee to indicate an answer to a question, such as a mark on an answer sheet, a handwritten essay, or an entry in an electronic storage medium.

Restriction of Range—A case in which the variance of scores in an analysis sample is lower than the variance of scores in the population from which the sample was selected. See *Analysis Sample*, *Variance*.

Sampling Error—The difference between a statistic derived from a particular sample and the corresponding parameter for the population from which the sample was drawn. See *Parameter* (1).

Score—A quantitative or categorical value (such as “pass” or “fail”) assigned to an examinee as the result of some measurement procedure.

Score Recipient—A person or institution obtaining the scores of individual examinees or summary data for groups of examinees.

Score Scale—The set of numbers within which scores are reported for a particular test or testing program, often, but not necessarily, having a specified mean and standard deviation for some defined reference group.

Special Testing Arrangement—A test administered under nonstandardized conditions in which modifications have been made to meet the needs of examinees who require the modifications for appropriate assessment, such as providing audiotaped versions of tests for visually impaired people. See *Standardized Conditions*.

Speededness—The extent to which test takers have time to respond to items on a test. One indicator of speededness is the percent of test takers who answer all of the items in the test.

Sponsors—Educational, professional, or occupational associations; federal, state or local agencies; public or private foundations that contract with ETS for its services. This term includes their governing boards, membership, and appointed committees or staff.

Standard Deviation—A statistic characterizing the magnitude of the differences among a set of measurements. Specifically it is the square root of the average squared difference between each measurement and the mean of the measurements. See *Variance*. The standard deviation is the square root of the variance.

Standard Error of Estimate—A statistic that indicates the standard deviation of differences between actual and estimated measures. It is an indicator of the accuracy of the estimate. See *Standard Deviation*.

Standard Error of Measurement—A statistic that indicates the standard deviation of the differences between observed scores and their corresponding true scores. It is also the standard deviation of scores for a person taking a large number of parallel forms of tests, assuming no changes in the person's true score. See *True Score*, *Standard Deviation*.

Standardized Conditions—The administration of a test in the same manner to all examinees to allow fair comparison of their scores. Factors such as timing, directions, and use of aids (e.g., calculators and dictionaries) are controlled to be constant for all examinees.

Standards for Educational and Psychological Testing, American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). Washington, D.C.: APA, 1985.

Subject—An individual who participates in an ETS laboratory or experimental research project.

Subscore—A score derived from a subset of the items in a test. Synonymous with *Part Score*.

Subtest—A subset of the items in a test upon which a subscore or part score is based.

Test Analysis—A description of the statistical characteristics of a test following administration, including but not limited to distributions of item difficulty and discrimination indices, score distributions, mean and standard deviation of scores, reliability, standard error of measurement, and indices of speededness.

Test Battery—(1) A collection of tests often administered together. (2) A collection of measures designed to allow the comparison of scores across measures for an individual.

Test Edition—A unique version of a test consisting of all of the identical copies of a test. Compare *Alternate Form*.

Test Format—The physical layout of a test, including the spacing of items on a page, type size, positioning of item-response options, etc.

Test Program Statistics—Data that are based on the groups of people that actually take the tests offered by a particular testing program. Program statistics are not equivalent to data derived from carefully selected samples of defined populations such as those used to construct national norms.

Testing Program—A comprehensive ongoing service under which examinees are scheduled to take a test under standardized conditions, the tests are supplied with instructions for giving and taking them, and arrangements are made for scoring the tests, reporting the scores, and providing interpretative information. A program is characterized by its continuing character and by the inclusiveness of the services provided.

Test-Retest Reliability—An estimate of reliability based on the correlation between scores on two administrations of the same test form to the same group of people. See *Reliability*. Compare *Alternate-Form Reliability*.

Test Specifications—Detailed documentation of the intended characteristics of a test, including but not limited to the content and skills to be measured, the number and type of items, level of difficulty and discrimination, the timing, and the layout.

Test-Taking Population (Intended)—The people for whom a test has been designed to be most appropriate. The actual test-taking population may differ in some instances from the intended population.

Timeliness—The degree to which a product or service is released or delivered to its recipient within a predefined schedule.

True Score—The hypothetical average score of an examinee calculated from an infinite number of administrations of alternate test forms, assuming no learning, forgetting, or fatigue on the part of the examinee. It is the score that an examinee would obtain if the test were perfectly reliable (the standard error of measurement were zero). See *Reliability, Standard Error of Measurement*.

Validity—The extent to which inferences and actions made on the basis of test scores are appropriate and justified by evidence.

Variance—A statistic characterizing the magnitude of the differences among a set of measurements. Specifically, it is the average squared difference between each measurement and the mean of the measurements.

Weighting System—(1) A formula giving the relative contribution of part scores to a composite score. See *Composite Score*. (2) The relative contribution assigned to certain sample data to better represent a population.

CODE OF FAIR TESTING PRACTICES IN EDUCATION

Prepared by the Joint Committee on Testing Practices

The Code of Fair Testing Practices in Education states the major obligations to test takers of professionals who develop or use educational tests. The Code is meant to apply broadly to the use of tests in education (admissions, educational assessment, educational diagnosis, and student placement). The Code is not designed to cover employment testing, licensure or certification testing, or other types of testing. Although the Code has relevance to many types of educational tests, it is directed primarily at professionally developed tests such as those sold by commercial test publishers or used in formally administered testing programs. The Code is not intended to

cover tests made by individual teachers for use in their own classrooms.

The Code addresses the roles of test developers and test users separately. Test users are people who select tests, commission test development services, or make decisions on the basis of test scores. Test developers are people who actually construct tests as well as those who set policies for particular testing programs. The roles may, of course, overlap as when a state education agency commissions test development services, sets policies that control the test development process, and makes decisions on the basis of the test scores.

The Code has been developed by the Joint Committee on Testing Practices, a cooperative effort of several professional organizations, that has as its aim the advancement, in the public interest, of the quality of testing practices. The Joint Committee was initiated by the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education. In addition to these three groups, the American Association for Counseling and Development/Association for Measurement and Evaluation in Counseling and Development, and the American

Speech-Language-Hearing Association are now also sponsors of the Joint Committee.

This is not copyrighted material. Reproduction and dissemination are encouraged. Please cite this document as follows:

Code of Fair Testing Practices in Education. (1988) Washington, D.C.: Joint Committee on Testing Practices. (Mailing Address: Joint Committee on Testing Practices, American Psychological Association, 750 First Avenue, NE, Washington, D.C., 20002-4242.)

Code of Fair Testing Practices in Education

The Code presents standards for educational test developers and users in four areas:

- A. Developing/Selecting Tests
- B. Interpreting Scores
- C. Striving for Fairness
- D. Informing Test Takers

Organizations, institutions, and individual professionals who endorse the Code commit themselves to safeguarding the rights of test takers by following the principles listed. The Code is intended to be consistent with the relevant parts of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985). However,

the Code differs from the Standards in both audience and purpose. The Code is meant to be understood by the general public; it is limited to educational tests; and the primary focus is on those issues that affect the proper use of tests. The Code is not meant to add new principles over and above those in the Standards or to change the meaning of the Standards. The goal is rather to represent the spirit of a selected portion of the Standards in a way that is meaningful to test takers and/or their parents or guardians. It is the hope of the Joint Committee that the Code will also be judged to be consistent with existing codes of conduct and standards of other professional groups who use educational tests.

A Developing/Selecting Appropriate Tests*

Test developers should provide the information that test users need to select appropriate tests.

Test Developers Should:

1. Define what each test measures and what the test should be used for. Describe the population(s) for which the test is appropriate.
2. Accurately represent the characteristics, usefulness, and limitations of tests for their intended purposes.
3. Explain relevant measurement concepts as necessary for clarity at the level of detail that is appropriate for the intended audience(s).
4. Describe the process of test development. Explain how the content and skills to be tested were selected.
5. Provide evidence that the test meets its intended purpose(s).
6. Provide either representative samples or complete copies of test questions, directions, answer sheets, manuals, and score reports to qualified users.
7. Indicate the nature of the evidence obtained concerning the appropriateness of each test for groups of different racial, ethnic, or linguistic backgrounds who are likely to be tested.
8. Identify and publish any specialized skills needed to administer each test and to interpret scores correctly.

Test users should select tests that meet the purpose for which they are to be used and that are appropriate for the intended test-taking populations.

Test Users Should:

1. First define the purpose for testing and the population to be tested. Then, select a test for that purpose and that population based on a thorough review of the available information.
2. Investigate potentially useful sources of information, in addition to test scores, to corroborate the information provided by tests.
3. Read the materials provided by test developers and avoid using tests for which unclear or incomplete information is provided.
4. Become familiar with how and when the test was developed and tried out.
5. Read independent evaluations of a test and of possible alternative measures. Look for evidence required to support the claims of test developers.
6. Examine specimen sets, disclosed tests or samples of questions, directions, answer sheets, manuals, and score reports before selecting a test.
7. Ascertain whether the test content and norms group(s) or comparison group(s) are appropriate for the intended test takers.
8. Select and use only those tests for which the skills needed to administer the test and interpret scores correctly are available.

*Many of the statements in the Code refer to the selection of existing tests. However, in customized testing programs test developers are engaged to construct new tests. In those situations, the

test development process should be designed to help ensure that the completed tests will be in compliance with the Code.

B Interpreting Scores

Test developers should help users interpret scores correctly.

Test Developers Should:

9. Provide timely and easily understood score reports that describe test performance clearly and accurately. Also explain the meaning and limitations of reported scores.
10. Describe the population(s) represented by any norms or comparison group(s), the dates the data were gathered, and the process used to select the samples of test takers.
11. Warn users to avoid specific, reasonably anticipated misuses of test scores.
12. Provide information that will help users follow reasonable procedures for setting passing scores when it is appropriate to use such scores with the test.
13. Provide information that will help users gather evidence to show that the test is meeting its intended purpose(s).

Test users should interpret scores correctly.

Test Users Should:

9. Obtain information about the scale used for reporting scores, the characteristics of any norms or comparison group(s), and the limitations of the scores.
10. Interpret scores taking into account any major differences between the norms or comparison groups and the actual test takers. Also take into account any differences in test administration practices or familiarity with the specific questions in the test.
11. Avoid using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use.
12. Explain how any passing scores were set and gather evidence to support the appropriateness of the scores.
13. Obtain evidence to help show that the test is meeting its intended purpose(s).

C Striving for Fairness

Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test Developers Should:

14. Review and revise test questions and related materials to avoid potentially insensitive content or language.
15. Investigate the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors.
16. When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores.

Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test Users Should:

14. Evaluate the procedures used by test developers to avoid potentially insensitive content or language.
15. Review the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available. Evaluate the extent to which performance differences may have been caused by inappropriate characteristics of the test.
16. When necessary and feasible, use appropriately modified forms of tests or administration procedures for test takers with handicapping conditions. Interpret standard norms with care in the light of the modifications that were made.

D Informing Test Takers

Under some circumstances, test developers have direct communication with test takers. Under other circumstances, test users communicate directly with test takers. Whichever group communicates directly with test takers should provide the information described below.

Test Developers or Test Users Should:

17. When a test is optional, provide test takers or their parents/guardians with information to help them judge whether the test should be taken, or if an available alternative to the test should be used.
18. Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Strive to make such information equally available to all test takers.

Under some circumstances, test developers have direct control of tests and test scores. Under other circumstances, test users have such control. Whichever group has direct control of tests and test scores should take the steps described below.

Test Developers or Test Users Should:

19. Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, retake tests, have tests rescored, or cancel scores.
20. Tell test takers or their parents/guardians how long scores will be kept on file and indicate to whom and under what circumstances test scores will or will not be released.
21. Describe the procedures that test takers or their parents/guardians may use to register complaints and have problems resolved.

Note: The membership of the Working Group that developed the Code of Fair Testing Practices in Education and of the Joint Committee on Testing Practices that guided the Working Group was as follows:

Theodore P. Bartell
John R. Bergan
Esther E. Diamond
Richard P. Duran
Lorraine D. Eyde
Raymond D. Fowler
John J. Fremer
(Co-chair, JCTP and Chair,
Code Working Group)

Edmund W. Gordon
Jo-Ida C. Hansen
James B. Lingwall
George F. Madaus
(Co-chair, JCTP)
Kevin L. Moreland
Jo-Ellen V. Perez
Robert J. Solomon
John T. Stewart

Carol Kehr Tittle
(Co-chair, JCTP)
Nicholas A. Vacc
Michael J. Zieky
Debra Boltas and Wayne
Camara of the American
Psychological Association
served as staff liaisons

Additional copies of the Code may be obtained from the National Council on Measurement in Education, 1230 Seventeenth Street, NW, Washington, D.C. 20036. Single copies are free.



Appendix G

Additional Issues Concerning the Use of Calculators on HSAs

What type of staff development will be required ?

Teachers, especially those who are required to use a graphing calculator, must receive training in the use of these relatively new pieces of equipment. Many math and science teachers, especially those teaching early high school courses, may have little or no experience in the use of graphing calculators. Still more teachers may need instructional materials and strategies for providing instruction to students in the use of graphing calculators. Teachers will need the equivalent of two days training in the use of graphing calculators and then forms of follow-up support will be required as teachers introduce these tools to students. This training should begin by the winter of 1998.

What are the potential implications of calculator use for student performance ?

While the use of graphing calculators is consistent with national curriculum standards, premature requirement of their use on high stakes assessments are likely to lead to extremely high failure rates and adverse impact against students in school settings who are slower to purchase and use advanced calculators in instruction. Many standardized tests now require or permit calculator use; however, relatively few assessments require graphing calculators—these functions are still generally reserved for assessments associated with third or fourth year math courses. As math reform progresses, students will be exposed to these skills at an early age and course level, but this migration of standards to entry-level math courses will likely require several more years in many districts and schools.

Even if instruction were offered by 1999-2000, students who had been using a graphing or scientific calculator in earlier courses are likely to be advantaged. Calculator use, like many educational reforms, will require several years for students, and teachers, to catch up. MSBE should be aware that math and science scores may be artificially low due to this effect and that results may be initially inequitable across districts (this effect could persist well into the use of the HSA for individual high-stakes uses). The TASC has asked that a pilot test be conducted prior to the state-wide field testing to determine student proficiency on items where graphing calculators are required and to determine if familiarity and type of calculator (functions, model) moderate performance. This pilot test would likely occur early in 1998-99. If the pilot test demonstrates that the vast majority of students are not proficient in the use of graphing calculators, the tests could become more a measure of calculator skills than the Core Learning Goals specified in the test specifications (i.e., if students lack a “basic skill” in calculator operations, they may be unable to demonstrate their skills on other math goals where a calculator is required). MSDE would then need to determine whether to: (a) continue to

mandate calculator skills on the tests (if this were done, students with proficiency in math could fail the test if they lack one skill—calculator use), or (b) transition the use of calculators over a few years until there are sufficient proportions of students demonstrating skill in the use of graphing calculators so that it would not interfere with test performance in math or science. As a first step, CB/ETS believe a survey of teachers should be conducted this fall to determine how many student currently use graphing calculator in math courses, how proficient they are in the use of a graphing calculator, what forms of staff development and support would be required to provide student instruction, etc.

What types or models of calculators will be used in courses ?

MSDE should specify the exact functions graphing calculators should possess as well as any functions (e.g., programmability, QWERTY keyboards, printing calculators) which might compromise test performance or security, and would be disallowed. While some of this work has been conducted, there is a need to align these requirements with specific models and to conduct an annual review to aide schools in purchases. The same process should be followed in each science area.

Who will purchase these calculators and where will funds come from?

The TASC recommends that state and local districts purchase the number of calculators required to complete each test. The number required for each test will be the total number of students in the school enrolled in a course (e.g., all Algebra I students, all Chemistry students), a figure several times larger than the number of calculators required for instruction. Typically, one teacher can purchase 8-12 graphing calculators that can be used by all students in several classes. However, if all students are to complete the HSA Algebra test at the same time, each student now will require the calculator.

What calculators should be purchased ?

As illustrated above, math wishes all students to have graphing calculators for consistency with the Core Learning Goals. Three areas of science prefer three different types of calculators. Because it would not be feasible in most schools for faculty to share calculators used for daily instruction, Maryland schools may have to purchase over 110,000 calculators (not including replacements) to initiate the HSA.¹

¹ The estimate of 110,000 does not include existing calculators in schools which may or may not be appropriate after the specifications teams detail the required functions. There are about 225,000 students in Maryland schools in grades 9-12. Assume 25 % of these students will be enrolled in Algebra 1 and another 25% enrolled in Geometry, this would require 113,000 graphing calculators for math alone. Science enrollments are much more difficult to estimate, but because each student must complete two science courses it is likely that 20% of students may enroll in Earth/Space Science, 15% in Chemistry, and less than 10% in Physics (because Biology does not require calculators for its test we have not added its enrollment). This would require about 20,000 additional graphing calculators for Physics, 34,000 scientific calculators for

There are two alternatives that would reduce the calculator expenditures for the state and local districts. First, if the same basic calculator were required across all four subjects, Maryland might only require enough calculators to: (a) provide each student with a calculator in the highest volume test during the HSA administrations (Algebra's estimated enrollment is approximately 58,000 students), and (b) provide each teacher with a sufficient number for instructional purposes (about .33 of their largest classroom, a number which would total less than the 58,000 calculators needed for the HSA). Under this arrangement, scientific calculators or graphing calculators would be selected as a common tool for all courses. Next, agreement must be reached about which course students would receive initial instruction in calculator use. Valuable instructional time might be devoted to calculator instruction at the cost of other educational goals and expectations. It is simply not educationally efficient to require math and science teachers to each devote substantial instructional time to graphing calculators. But this would be required if differences in course sequences remain and use of a graphing calculator is required for successful performance on these HSA tests. This would be a significant issue if graphing calculators were selected because substantial practice and instruction may be required for students to become familiar with the functions in order to be successful in these courses. It appears easier for Algebra I to serve as a prerequisite course.

A second alternative would be to purchase a graphing calculator for each student as they enter Algebra I. This would be the same calculator across the state and each year MSDE would determine which model would be purchased. The calculator would be given to students free, but if it were lost, stolen or broken the student would need to reimburse the school for the purchase of a replacement calculator (not unlike local policies regarding some textbooks). In this way, the student would have a calculator for 4 years of school. Each year about 58,000 calculators may need to be purchased, and this cost would be part of the fixed costs associated with HSA. MSDE should also consider introducing calculator instruction and use during middle school and prior to Algebra courses.

Under all these scenarios, MSDE would need to select potential models or the one model that will serve all content areas equally. The state would also realize a significant cost increase if one large purchase were made by the state rather than a series of smaller purchases from local districts or schools. Finally, because technology changes so quickly in this area, MSDE should anticipate the need to replenish about 25% of the total calculators each year—with this assumption the latter two alternatives do not appear as unreasonable as they may have seemed initially.

Chemistry, and 45,000 4-function calculators. This is a significant purchase for the state and districts without replacements considered.

The TASC has recommended that a committee be formed this summer to develop a proposal to support a request from the General Assembly for the necessary equipment for the HSA. The TASC was unable to identify additional equipment needed for the HSA because individual Test Specification Committees are still working on this task. The lack of closure on this issue and the need to acquire the equipment and train staff and students in its use presents a very serious threat to the present HSA schedule. If the equipment and training are not in place by September 1998, state-wide field testing cannot accomplish all the necessary objectives, and operational implementation of the HSA may need to be deferred.

PRINTED BY THE DEPARTMENT OF GENERAL SERVICES - VISUAL COMMUNICATIONS AND DIGITAL IMAGING - 410-767-4594
ON RECYCLED PAPER

